(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
23 August 2001 (23.08.2001)

PCT

(10) International Publication Number
WO 01/61036 A2

(51) International Patent Classification[7]: C12Q 1/68

(21) International Application Number: PCT/GB01/00718

(22) International Filing Date: 19 February 2001 (19.02.2001)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:

20000792	17 February 2000 (17.02.2000)	NO
20012864	21 February 2000 (21.02.2000)	NO
20012863	27 February 2000 (27.02.2000)	NO

(71) Applicant *(for all designated States except US)*: **COMPLETE GENOMICS AS** [NO/NO]; P.O. Box 64, Blindern, N-0313 Oslo (NO).

(71) Applicant *(for MW only)*: **TOWLER, Philip, Dean** [GB/GB]; Frank B. Dehn & Co., 179 Queen Victoria Street, London EC4V 4EL (GB).

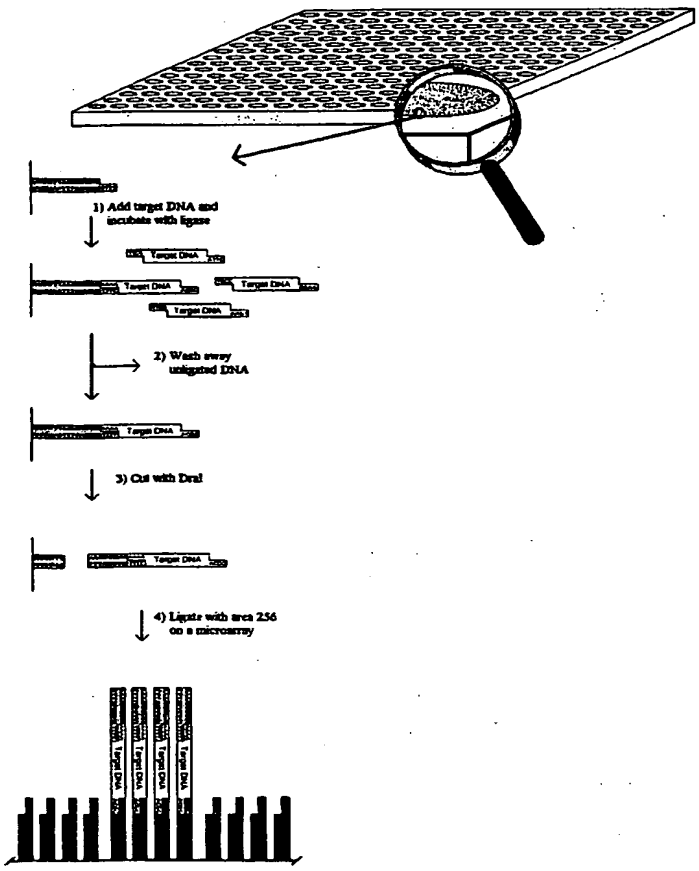
(72) Inventor; and
(75) Inventor/Applicant *(for US only)*: **LEXOW, Preben** [NO/NO]; Bloksbergveien 16, N-3132 Husøysund (NO).

(74) Agents: **TOWLER, Philip, Dean** et al.; Frank B. Dehn & Co., 179 Queen Victoria Street, London EC4V 4EL (GB).

(81) Designated States *(national)*: AE, AG, AL, AM, AT, AT (utility model), AU, AZ, BA, BB, BG, BR, BY, BZ, CA, CH, CN, CR, CU, CZ, CZ (utility model), DE, DE (utility model), DK, DK (utility model), DM, DZ, EE, EE (utility model), ES, FI, FI (utility model), GB, GD, GE, GH, GM, HR, HU, ID, IL, IN, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MA, MD, MG, MK, MN, MW, MX, MZ, NO, NZ, PL, PT, RO, RU, SD, SE, SG, SI, SK, SK

[Continued on next page]

(54) Title: A METHOD OF MAPPING RESTRICTION ENDONUCLEASE CLEAVAGE SITES



(57) Abstract: The invention provides a two-step sorting procedure where it is possible to scan the overhanging single-stranded ends of nucleic acid fragments quickly and efficiently using solid supports, such as microarrays. Use is made of two different sets of degenerate overhang-adaptors in this regard. The invention also provides new methods and strategies inter alia for collecting information about sequences and cleavage sites that are between the cleavage sites that have generated an overhang pair. An effective method of producing the restriction map, making it easier to create multiple maps, is also described.

(utility model), SL, TJ, TM, TR, TT, TZ, UA, UG, US, UZ, VN, YU, ZA, ZW.

(84) **Designated States** *(regional)*: ARIPO patent (GH, GM, KE, LS, MW, MZ, SD, SL, SZ, TZ, UG, ZW), Eurasian patent (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European patent (AT, BE, CH, CY, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE, TR), OAPI patent (BF, BJ, CF, CG, CI, CM, GA, GN, GW, ML, MR, NE, SN, TD, TG).

Published:

— *without international search report and to be republished upon receipt of that report*

For two-letter codes and other abbreviations, refer to the "Guidance Notes on Codes and Abbreviations" appearing at the beginning of each regular issue of the PCT Gazette.

A Method of Mapping Restriction Endonuclease Cleavage Sites

The present invention relates to new methods of method of mapping restriction endonuclease cleavage sites.

Traditionally, DNA molecules have been mapped using Type II restriction endonucleases such as EcoRI and HindIII which have well-defined recognition and cleavage sites. After cleavage with the restriction endonucleases, the DNA fragments are generally run on an agarose gel together with DNA markers of known size and visualised using EtBr under UV light.

More recently, use has been made of Type IIs restriction endonucleases which have cleavage sites outside their recognition sites. Reference is made in this regard to US 5,858,656 and Gene, 145 (1994) 163-169.

However, there remains a need to provide effective methods for determining the sequence of the single-strand overhangs that are created with Type IIs restriction endonucleases. The invention therefore provides a two-step sorting procedure where it is possible to scan the overhangs quickly and efficiently using solid supports such as microarrays. Furthermore, the invention provides new methods and strategies inter alia for collecting information about sequences and cleavage sites that are between the cleavage sites that have generated an overhang pair. An effective method of producing the restriction map, making it easier to create multiple maps, is also described.

The invention therefore provides a method of mapping a target nucleic acid molecule, the method comprising the steps of:

(a) treating the target nucleic acid molecule with one or more restriction endonucleases to produce one or more nucleic acid fragments having first and second 5'- or 3'- single-stranded overhanging ends,

(b) adding the nucleic acid fragments to a first set of overhang-adaptors,

each overhang-adaptor of the first set comprising a nucleic acid molecule comprising at least one 5'- or 3'-single-stranded end,

the single-stranded ends of the overhang-adaptors being of lengths and orientations (i.e. 5'- or 3'-) corresponding to the lengths and orientations of the overhanging single-strands of the cleavage sites of the said restriction endonucleases,

wherein said first set comprises a collection of overhang-adaptors whose single-stranded ends collectively encode up to all possible permutations and combinations of the nucleotides A, C, G and T,

and wherein each overhang-adaptor in the said first set is spatially separable from every other different overhang-adaptor in the first set;

(c) contacting the said nucleic acid fragments with a nucleic acid ligase to cause selective ligation of the nucleic acid fragments with those overhang-adaptors whose 5'- or 3'- single-stranded ends are fully complementary to the 5'- or 3'-overhanging single-stranded ends of the nucleic acid fragments,

thus forming a plurality of separable populations of nucleic acid fragments which are ligated at their first ends to a first overhang-adaptor;

optionally, removing the unligated nucleic acid fragments;

(d) identifying the sequence of the second overhanging single-stranded end of the nucleic acid fragments; and

(e) comparing the sequences of the ends of the nucleic acid fragments in order to produce a map of the target nucleic acid molecule.

The invention also provides methods for identifying the overhanging ends of a nucleic acid fragment comprising the steps (b)-(d) as described above.

As used herein, the term "mapping a target nucleic acid molecule" means providing information on the order of some or all of the fragments into which the target nucleic acid molecule may be divided or on the position of discrete sequences, e.g. restriction endonuclease cleavage sites, within the target nucleic acid molecule. The mapping of a target nucleic acid molecule

will often facilitate its subsequent sequencing.

As used herein the term "target nucleic acid molecule" refers to any nucleic acid molecule, for example a naturally occurring, synthetic or recombinant polynucleotide molecule. The term includes DNA, such as genomic, cDNA or vector DNA; RNA, such as mRNA; and PNA and their analogues. Generally, the term relates to a double-stranded nucleic acid molecule, most preferably a DNA molecule.

The target nucleic acid molecule is treated, i.e. digested, cleaved or cut, with one or more restriction endonucleases in order to divide up the target nucleic acid molecule into one or more nucleic acid fragments. Each of these fragments will have two ends, i.e. the first and second ends, having overhanging, i.e. single-stranded, stretches of nucleotides.

The invention particularly relates to the use of restriction endonucleases which cleave DNA to produce overhanging ends which are non-identical in sequence and/or have overhanging sequences which are unrelated to the recognition sequence of the restriction enzyme used. Preferably, the restriction endonuclease is a Type Ip or Type IIs restriction endonuclease.

Type Ip restriction endonucleases generate degenerate overhangs in the middle of their recognition sequences.

Type IIs restriction endonucleases interact with two discrete sites on double-stranded DNA: the recognition site which is 4-7bp (bp=base pairs) long and the cleavage site which is usually 1-20bp away from the recognition site. Overhangs of -6 to +5 nucleotides are usually produced. These endonucleases exhibit no specificity to the sequence that is cut and they can therefore generate overhangs with all types of nucleotide compositions.

Over 70 classes of Type IIs restriction endonucleases have been identified and there are large variations both with respect to substrate specificity and cleavage pattern. In addition, these enzymes have proved to be well suited to "module swapping" experiments so that one can create new enzymes for particular requirements (Huang-B, et al.; J-Protein-Chem. 1996, 15(5):481-9, Bickle, T.A.; 1993 in Nucleases (2nd edn), Kim-YG et al.; PNAS 1994, 91:883-887). Very many combinations and variants of these enzymes can therefore be used according to the principles described herein.

Examples of Type IIs restriction endonucleases which may be used in this regard include Bbv I, Bce83 I, Bcef I, Bmp I, Bsg I, BspLU11 III, Bst71 I, Eco57 I, Fok I, Gsu I, Hga I, Mme I and the like.

Preferably, Type IIs restriction endonucleases are used which produce

overhangs of 3-5 nucleotides, preferably 3 or 4 nucleotides, either at the 5'-end or the 3'-end of the nucleic acid fragment.

Particularly preferred restriction endonucleases are AlwNI, Bs1I, DraIII, PflMI, BstXI, BplI, BaeI, EarI, SapI, BbsI, BbvI, BsaI, FokI, SfaNI and HgaI.

In one preferred embodiment of the invention, combinations of Type IIs restriction endonucleases are used which either all produce 5'-overhangs or all produce 3'-overhangs. This obviates the need for sets of overhang-adaptors with both 5'- and 3'-single-stranded ends.

Alternatively, the restriction endonuclease is one with an interrupted palindromic recognition sequence which cuts at sites which are independent of the intervening sequences, provided that the intervening sequence is of the appropriate length.

In the context of this invention, any reference to a Type IIs restriction endonuclease should also be considered to be a reference to a Type Ip restriction endonuclease.

In a preferred embodiment of the invention, the target nucleic acid molecule is treated with only one restriction endonuclease. In this case, the restriction endonuclease is preferably a Type Ip or IIs restriction endonuclease.

In another preferred embodiment of the invention, the target nucleic acid molecule is treated with more than one restriction endonuclease, wherein the restriction endonucleases either all produce 5'-overhanging ends or all produce 3'-overhanging ends.

The digested nucleic acid fragments are then added to a first set of overhang-adaptors.

In the context of the present invention, the term "overhang-adaptor" refers to a structure comprising a nucleic acid molecule comprising, i.e. consisting at least of, a 5'- or 3'- single-stranded nucleic acid end.

The essential feature of each of the overhang-adaptors is that they possess at least one free 5'- or one free 3'- single-stranded nucleic acid end. The remaining part(s) of the overhang adaptor should allow the binding of the single-stranded end of the overhang-adaptor to a single-stranded end of the nucleic acid fragments. For example, the remaining part of the overhang-adaptor may be a single-stranded or double-stranded nucleic acid molecule, preferably a DNA molecule. Most preferably, the overhang-adaptor is a single-stranded DNA molecule or oligonucleotide.

In this context, the term "single-stranded ends" of the overhang-

adaptors refers to that part of the overhang-adaptor which might be complementary to a single-stranded overhang of the nucleic acid fragments.

Thus it can be seen that the end of the overhang-adaptor which binds to the nucleic acid fragment may be single-stranded DNA and also the remaining part of the overhang-adaptor may be single-stranded DNA. The single-stranded DNA may, for example, be an oligonucleotide of total length 10-50 nucleotides, preferably 12-30 nucleotides, and most preferably 13-20 nucleotides. In some embodiments of the invention, overhang-adaptors which are double-stranded DNA molecules having single-stranded 5'- or 3'-overhangs are excluded.

The single-stranded ends of the overhang-adaptors are of lengths and orientations which correspond to the lengths and orientations of the overhanging single-strands of the cleavage sites of the restriction endonucleases used.

The lengths and orientations of the cleavage sites of the restriction endonucleases will be known in each case. In this context, the term "orientation" merely refers to whether the single-stranded overhang produced by cleavage with the restriction endonuclease is a 5'-overhang or a 3'-overhang.

It will be appreciated that the single-stranded ends of the nucleic acid fragments and overhang-adaptors must be generally complementary in form, i.e. where the nucleic acid fragments all have 5'- single-stranded overhangs, the single-stranded ends of the overhang-adaptors (both the first and second sets) will all be 5'- to allow binding thereto; and where the nucleic acid fragments all have 3'- single-stranded overhangs, the single-stranded ends of the overhang-adaptors (both the first and second sets) will all be 3'- to allow binding thereto. Where the nucleic acid fragments have combinations of 5'- and 3'-single-stranded overhangs, then the sets of the overhang-adaptors must also contain adaptors having 5'- and 3'-single-stranded ends.

In one embodiment of the invention, the overhang-adaptors are single-stranded DNA molecules which have mirror-image sequences at each end (for example, 5'-CATC---------GTAG-3'). In between the sequences is a stretch of DNA or other structure which allows the overhang-adaptor to form a loop. The overhang-adaptor is then bound to a solid support, if necessary, in the region between the two end sequences. In this way, overhang-adaptors at any one spatial location or address will bind to the same specific single-stranded sequence whether that sequence is a 5'-single-stranded sequence or a 3'-single-stranded sequence.

With regard to any one restriction endonuclease that is used, it will be appreciated that the single-stranded ends of the nucleic acid fragments and the single-stranded ends of the overhang-adaptors must be generally of the same length. Thus for example, where the nucleic acid fragments all have 5'-single-stranded overhangs of length n, the single-stranded nucleic acids of the overhang-adaptors (both the first and second sets) will all be 5'- single-stranded overhangs of length n to allow binding thereto; and where the nucleic acid fragments all have 3'- single-stranded overhangs of length n, the single-stranded nucleic acids of the overhang-adaptors (both the first and second sets) will all be 3'- single-stranded overhangs of length n to allow binding thereto.

However, if the chosen combination of restriction endonucleases produces overhangs of different lengths, then the set of overhang adaptors will need to comprise single-stranded ends which are capable of binding to each of these different length overhangs. It will be appreciated, however, that if adaptors are used having single-stranded ends of a length that corresponds to the longest of the overhangs produced by the chosen restriction endonucleases, then the ends of such adaptors should also be capable of binding the shorter overhangs. Under such circumstances, a modification of the method used to identify the nucleic acid fragments which have been ligated to the second overhang-adaptor might be required.

The first set comprises a collection of overhang-adaptors whose 5'- and/or 3'- single-stranded ends collectively encode up to all possible permutations and combinations of the nucleotides A, C, G and T, i.e. the single-stranded ends comprise a set of degenerate sequences of nucleotides corresponding to the length and orientation of the overhanging ends of the nucleic acid fragments. Thus within the set of overhang-adaptors, there will be individual overhang-adaptors that are capable of hybridising and ligating to each of the individual first ends of the nucleic acid fragments. In some embodiments, universal nucleotides may be used at one or more of the positions in the single-stranded ends of the overhang adaptors.

For example, if the length of the overhang produced by the restriction endonuclease is 4, then the first set of overhang-adaptors will comprise AAAA, AAAC, AAAG, AAAT, AACA, AACC, etc.. In general, where n is the length of the overhang, the first set of overhang adaptors will consist of all or essentially all of 4^n adaptors. Thus where n=4, a set of 256 overhang-adaptors will be used.

If combinations of restriction endonucleases are used, all of which

produce overhangs of the same orientation and of the same length n, then generally a set of 4^n overhang-adaptors will be required. However, if combinations of restriction endonucleases are used all of which produce overhangs of the same length n but with different orientations, then generally a set of 2×4^n adaptors will be required. If restriction endonucleases are used which produce overhangs of different lengths, then the same principles apply, mutatis mutandis.

If desired, one or more of A, C, G or T may be replaced by an alternative nucleotide, i.e. U for T, or I. In particular, universal nucleotides which bind to A, C, G and T may be used in one or more positions in the overhang-adaptors.

It should be noted that the number of adaptors required may be reduced if not all of the nucleotides in an overhang are read. Thus it is possible to read only 3 out of 5 nucleotides in an overhang, thus reducing the number of required adaptors from 1024 to 64. In such a case, universal nucleotides which bind to A, C, G or T may be used in the adaptors.

For the purposes of the invention, each overhang-adaptor in the said first set will be spatially separable or spatially separate from every other different overhang-adaptor in the first set; and the spatial position or address of each overhang-adaptor and the sequence of its single-stranded end will be known. Thus, for example, it will be possible to distinguish between overhang-adaptors having AAAA single-stranded ends from overhang-adaptors having AAAC or AAAG single-stranded ends. In this context, therefore, the term "spatially separable" is intended to mean that the different overhang-adaptors might be spatially separated or physically separated from one another, for example, in separate compartments or wells, or attached to distinct or defined areas of a solid support, such as a microarray. In one embodiment of the invention, samples of each of the different overhang-adaptors of the first set are transferred for use in the second stage of the mapping method and hence each of the different overhang-adaptors needs to be physically distinguishable from all of the others.

After the nucleic acid fragments are added to the first set of overhang-adaptors, the nucleic acid fragments are contacted with a nucleic acid ligase to cause selective ligation of the nucleic acid fragments with those overhang-adaptors of the first set whose 5'- or 3'- single-stranded ends are fully complementary to the 5'- or 3'-overhanging single-stranded ends of the nucleic acid fragments. In this way, a plurality of separable or physically distinguishable populations of nucleic acid fragments are formed which are

ligated at their first ends to a first overhang-adaptor.

Preferably, the overhang-adaptors (of both sets) are treated with phosphatase prior to use in order to reduce the occurrence of ligation between adjacent overhang-adaptors.

Following the addition of nucleic acid ligase (which is preferably a DNA ligase), ligation is allowed to occur for an appropriate length of time for the single-stranded ends of the overhang-adaptors which are fully complementary to the overhanging single-stranded ends of the nucleic acid fragments to be ligated thereto.

The ligation step may be replaced by any other process which selectively binds the single-stranded ends of the overhang-adaptors to the fully complementary overhanging ends of the nucleic acid fragments.

In some embodiments of the invention, the reference to ligation and ligating the nucleic acid fragments may be replaced by a chemical ligation, such as that described in Nature Biotechnology, vol.19, February 2001, pp148-152, Xu et al.

Thus upon contacting the nucleic acid fragments with the first set of overhang-adaptors, the complementary ends of these two groups of molecules are allowed to hybridise and be ligated to one another. For example, if the target nucleic acid molecule is cut with Type IIs restriction endonuclease Fok I, 4-nucleotide 5'-overhanging ends will be produced in the nucleic acid fragments (assuming that at least one Fok I site is present in the target DNA). This might, for example, produce a 5'-overhanging end having the sequence 5'-GATC-3'. This overhanging end would then selectively hybridise to the overhang-adaptor with the 5'-end sequence of 5'-GATC-3'. Upon the addition of DNA ligase, the adjacent 3'-end of the nucleic acid fragment would then be ligated to the 5'-end of the overhang-adaptor.

The overhang-adaptors may either be attached to or carrying a means for attaching to a solid support.

In one preferred embodiment of the invention, overhang-adaptors are fixed to solid supports. This may be achieved in a number of different ways. The overhang-adaptors may be attached to one or more moieties which allow binding of that overhang-adaptor to a solid support, for example the end (or several internal sites) may be provided with one partner of a binding pair, e.g. with biotin which can then be attached to a streptavidin-carrying solid support.

Overhang-adaptors may be engineered to carry such a binding moiety in a number of known ways. For example, a PCR reaction may be conducted

to introduce the binding moiety, e.g. by using an appropriately-labelled primer. Alternatively, the overhang-adaptor may be ligated to a binding moiety, e.g. by cleaving the overhang-adaptor with a restriction enzyme and then ligating it to an adapter/linker whose end has been labelled with a binding moiety. Such a strategy would be particularly suitable if a Type IIs restriction endonuclease is used that forms a non-palindromic overhang. Another alternative is to clone the overhang-adaptor into a vector which already carries a binding moiety, or that contains sequences that facilitate the introduction of such a moiety.

Alternatively overhang-adaptors may be attached to solid supports without the need to attach a binding moiety insofar as the overhang-adaptor itself is one partner of the binding pair. Thus, for example short PNA molecules that are attached to a solid support may be used. PNA molecules have the ability to hybridize and bind to double-stranded DNA and overhang-adaptors can therefore be attached to a solid support with this strategy. Similarly, oligonucleotide probes may be used to bind complementary sequences to a solid support.

Appropriate solid supports suitable as immobilizing moieties for attaching the overhang-adaptors are well known in the art and widely described in the literature. Generally speaking, the solid support may be any of the well-known supports or matrices which are currently widely used or proposed for immobilization, separation, etc., in chemical or biochemical procedures. Thus for example, the immobilizing moieties may take the form of beads, particles, sheets, gels, wells, filters, membranes, microfibre strips, tubes or plates, fibres or capillaries, made for example of a polymeric material, e.g. agarose, cellulose, alginate, teflon, latex or polystyrene. Particulate materials, e.g. beads, are generally preferred. Conveniently, the immobilizing moiety may comprise magnetic particles, such as superparamagnetic particles. In a further preferred embodiment, plates or sheets are used to allow fixation of molecules in linear arrangement. The plates may also comprise walls perpendicular to the plate on which molecules may be attached. Attachment to the solid support may be performed directly or indirectly. For attaching the target molecules, conveniently attachment may be performed indirectly by the use of an attachment moiety carried on the nucleic acid molecules and/or solid support. Thus for example, a pair of affinity binding partners may be used, such as avidin, streptavidin or biotin, DNA or DNA binding protein (e.g. either the lac I repressor protein or the lac operator sequence to which it binds), antibodies (which may be mono- or polyclonal), antibody fragments

or the epitopes or haptens of antibodies. In these cases, one partner of the binding pair is attached to (or is inherently part of) the solid support and the other partner is attached to (or is inherently part of) the nucleic acid molecules. Other techniques of direct attachment may be used such as for example if a filter is used, attachment may be performed by UV-induced crosslinking. When attaching DNA fragments, the natural propensity of DNA to adhere to glass may also be used.

Attachment of appropriate functional groups to the solid support may be performed by methods well known in the art, which include for example, attachment through hydroxyl, carboxyl, aldehyde or amino groups which may be provided by treating the solid support to provide suitable surface coatings. Attachment of appropriate functional groups to the nucleic acid molecules of the invention may be performed by ligation or introduced during synthesis or amplification, for example using primers carrying an appropriate moiety, such as biotin or a particular sequence for capture.

Attachment to a solid support may be performed before or after overhang-adaptors have been produced. For example, overhang-adaptors carrying binding moieties may be attached to a solid support and thereafter treated with DNAse I or similar. Alternatively cleavage may be effected and then the fragments may be attached to the support.

Thus one strategy which may be used is to fix polynucleotides that complement the overhang-adaptors that are to be isolated to a solid support (the inside of a well, mono-dispersed spheres, microarrays, etc.).

Most preferably, the ligation reaction is carried out in free solution, i.e. where the overhang-adaptors are not attached to a solid support. The efficiency of the ligation reaction may be improved in this way. In such circumstances, the overhang-adaptors carry a means for attaching to a solid support, for example, biotin. Optionally, after the ligation reaction, the overhang-adaptors are bound to a solid support.

An alternative is to fix the overhang-adaptors to a solid support such as paramagnetic beads or similar.

A washing step is preferably carried out following the ligation of the nucleic acid fragments and the first set of overhang-adaptors in order to remove unligated nucleic acid fragments. The overhang-adaptors will generally be immobilised or bound to a solid support during the washing step.

It should be pointed out that the specificity of the method can be adjusted to most purposes by repeating steps (b) and (c) one or several times, with a washing step in between if desired. It may also be appropriate to use

competing probes/overhangs during step (b) in order to increase specificity.

At the end of step (c), a plurality of spatially separable or separate populations of nucleic acid fragments is formed which are ligated at their first ends to a first overhang-adaptor. Since the spatial position (i.e. the address) and the single-stranded end sequence of each of the first overhang-adaptors will be known, this will provide information on the sequences of the first overhanging ends of the nucleic acid fragments. Thus the sequence of the first overhanging end of each of the nucleic acid fragments is informationally linked to its spatial position or address.

It should be noted that the first overhanging ends of the nucleic acid molecules will at this point have been inactivated through ligation to the first overhang-adaptors, i.e. the first overhanging ends of the nucleic acid fragments will no longer be capable of binding to further overhang-adaptors. The second overhanging ends of the nucleic acid fragments will essentially still be unbound.

The ligation (and subsequent washing step, if required) marks the end of the first stage of the mapping method.

The sequences of the second overhanging single-stranded ends of the nucleic acid fragments are then identified. This may be done by a number of different ways:

Second stage - Method 1. Preferably, step (d) is carried out by:

(d1) optionally releasing each population of ligated nucleic acid fragments from the solid support,

selectively contacting each population of nucleic acid fragments which were ligated at their first ends to a first overhang-adaptor with a second set of overhang-adaptors,

each overhang-adaptor of the second set comprising a nucleic acid molecule comprising at least one 5'- or 3'-single-stranded end,

the single-stranded ends of the overhang-adaptors of the second set being of lengths and orientations (i.e. 5'- or 3'-) corresponding to the lengths and orientations of the overhanging single-strands of the cleavage sites of the said restriction endonucleases,

wherein said second set comprises a collection of overhang-adaptors

whose single-stranded ends collectively encode up to all possible permutations and combinations of the nucleotides A, C, G and T,

and wherein each overhang-adaptor in the said second set is spatially distinguishable from every other different overhang-adaptor in the second set;

(d2) contacting the nucleic acid fragments with a nucleic acid ligase to cause selective ligation of the nucleic acid fragments with those overhang-adaptors of the second set whose 5'- or 3'- single-stranded ends are fully complementary to the second 5'- or 3'-overhanging ends of the nucleic acid fragments;

thus forming a plurality of populations of nucleic acid fragments which are ligated at their second ends to a second overhang-adaptor, and

optionally removing the non-ligated nucleic acid fragments;

(d3) identifying the sequences of the first and second overhanging ends of each of the nucleic acid fragments from the spatial positions of the second overhang-adaptors to which the nucleic acid fragments are ligated.

In a preferred aspect of the invention, steps (b)-(d2) are carried out simultaneously, i.e. the nucleic acid fragments are combined with the first and second sets of overhang adaptors simultaneously with the nucleic acid ligase.

After the end of the first stage in the mapping procedure, the nucleic acid fragments are then prepared for contacting with the second set of overhang-adaptors. If the first overhang-adaptors were bound to a solid support, they are now released from that support, thus facilitating the transfer of the nucleic acid fragments to a different spatial position. The method of separation of the nucleic acid fragments from the solid support will be dependent on the way that the nucleic acid fragments were bound. One example of a method of releasing the nucleic acid fragments is through the use of a cleavage site located in the first overhang-adaptor. If the first overhang-adaptors are DNA molecules, then a restriction endonuclease that recognises a site in the (non-variable end of the) first overhang-adaptor may be used. Restriction endonucleases that produce overhanging ends having a length and

orientation which correspond to any of the second ends of the nucleic acid fragments should be avoided. Provided that the latter issue is taken into consideration, the nucleic acid fragments may be released through cleavage within the nucleic acid fragment itself.

Each individual population of nucleic acid fragments which were ligated at their first ends to a first overhang-adaptor is then selectively contacted with a second set of overhang-adaptors, i.e. each nucleic acid fragment population is independently contacted with a second set of overhang-adaptors. Thus for example, the population of nucleic acid fragments which were bound to first overhang-adaptors having the first end sequence AAAC will be contacted independently with the second set of overhang-adaptors compared to the population of nucleic acid fragments which were bound to first overhang-adaptors having the first end sequence AAAT. In this way, the positional information which was derived from the first stage of the mapping method is preserved.

The second set of overhang-adaptors are similar in many ways to those of the first step, particularly in the combinatorial nature of their single-stranded end sequences. Hence most of the comments given above regarding the first set of overhang-adaptors apply to the second set of overhang-adaptors, mutatis mutandis.

Thus each overhang-adaptor of the second set comprises a nucleic acid molecule comprising at least one 5'- or 3'- single-stranded end.

In the same manner as the first set of overhang-adaptors, the 5'- and/or 3'-single-stranded ends of the overhang-adaptors of the second set have lengths and orientations (i.e. 5'- or 3'-) corresponding to the lengths and orientations of the overhanging single-strands of the cleavage sites of the chosen restriction endonucleases, wherein the second set comprises a collection of overhang-adaptors whose single-stranded ends collectively encode up to all possible permutations and combinations of the nucleotides A, C, G and T.

Thus within the second set of overhang-adaptors, there will be individual overhang-adaptors that are capable of hybridising and ligating to each of the individual second ends of the nucleic acid fragments.

Each overhang-adaptor in the said second set is spatially separable or spatially identifiable from every other different overhang-adaptor in the second set. Thus the position of each different overhang-adaptor will be known and this positional information can be used to determine the sequence of the first and second overhanging ends of any nucleic acid fragment which is

bound thereto.

The second overhang-adaptors are preferably bound to a solid support, such as those described above. Most preferably, the solid support is a microarray, ideally one which can be automatically read, for example by a scanner.

After the populations of nucleic acid fragments are contacted with the second set of overhang-adaptors, they are then contacted with a nucleic acid ligase to cause selective ligation of the nucleic acid fragments with those overhang-adaptors of the second set whose 5'- or 3'- single-stranded ends are fully complementary to the second 5'- or 3'-overhanging ends of the nucleic acid fragments. The nucleic acid ligase is preferably a DNA ligase.

Preferably, the overhang-adaptors of the second set are treated with phosphatase prior to use in order to reduce the occurrence of ligation between adjacent overhang-adaptors.

Following the addition of nucleic acid ligase, ligation is allowed to occur for an appropriate length of time for the single-stranded ends of the second overhang-adaptors which are fully complementary to the second overhanging ends of the nucleic acid fragments to be ligated thereto. In this way, a plurality of populations of nucleic acid fragments which are ligated at their second ends to a second overhang-adaptor are formed.

The spatial positions of each of the second overhang-adaptors are positionally correlated with the sequences of the first and second ends of the nucleic acid fragments. Consequently, the identification of which of the second overhang-adaptors have nucleic acid fragments ligated thereto will provide information on sequences of the ends of all of the nucleic acid fragments, thus facilitating the mapping of the target nucleic acid molecule.

The following method may be used to determine which nucleic acid fragments have bound to the second overhang-adaptors.

The invention therefore also provides a method suitable for detecting overhangs on a microarray address, the method comprising the steps of:

providing one or more single-stranded nucleic acid adaptors each comprising a first part and a second part, the first and second parts being contiguous with one another, the first part having a free 5'- or 3'- end;

wherein the adaptor is preferably bound to a solid support;

contacting the adaptor with a target nucleic acid molecule having a single-stranded overhang which is complementary with the first part of the adaptor;

ligating the first part of the adaptor to the single-stranded overhang of the target nucleic acid molecule;

contacting the second part of the adaptor with one or more labelled single-stranded nucleic acid probes having a nucleotide sequence which is complementary with the second part of the adaptor;

ligating the labelled single-stranded nucleic acid probe to the target nucleic acid molecule;

optionally removing any unligated labelled single-stranded nucleic acid probe and/or unligated nucleic acid molecule;

determining whether any target nucleic acid molecule has been ligated to the first part of the adaptor by determining whether any labelled probe is bound to the second part of the adaptor.

It will be appreciated that the "one or more single-stranded nucleic acid adaptors" may comprise a first set of adaptors such as those defined above. Thus in one embodiment, the adaptors form a set of adaptors, the first parts of which are of lengths and orientations which correspond to the lengths and orientations of the overhanging single-stranded ends of the target nucleic acid molecules, wherein said set comprises a collection of adaptors whose first parts collectively encode up to all possible permutations and combinations of the nucleotides A, C, G and T, and wherein each adaptor in the said set is preferably spatially separable from every other different overhang-adaptor in the set. The comments given above with regard to first set of adaptors apply herein, mutatis mutandis.

The solid support will preferably be an array or microarray.

The labelled single stranded nucleic acid probes have nucleotide sequences which are complementary with all or essentially all of the second part of the adaptor, such that they are capable of hybridising to the adaptor and ligating with a target nucleic acid molecule when such a target nucleic acid molecule is bound by the first part of the adaptor.

The ligation steps may be carried out either sequentially or simultaneously. In the latter case, the target nucleic acid molecule is contacted with the adaptor together with the probe, and ligase is then added.

In most embodiments of the invention, the ligation steps will be carried out as described above, i.e. sequentially. This allows competing non-labelled probes to be used to reduce the background levels of the method.

The nucleic acid adaptor is preferably DNA. Similarly, the probe is also preferably DNA. The ligase is preferably a DNA ligase. The probe is preferably labelled with a fluorescent moiety.

It will be appreciated that for the probe to be ligatable to the target nucleic acid molecule, one end of the probe must be capable of hybridising to the adaptor at a position such that the ends of the target nucleic acid molecule and probe are contiguous.

The following is an example of this method. In this example, an overhang of the 4 bases TGCA is to be registered. (The principle is the same for 3`- and 5`- overhangs). This example is illustrated in Figure 1.

Oligonucleotides with the sequence GCGGATGCAGGACGT attached to a microarray are the basis for this example. The first (innermost) 11 nucleotides are designed to complement a fluorescent probe, while the 4 last (outermost) nucleotides will recognize the overhang of the target nucleic acid molecule. There is evidently great freedom of choice regarding the length and arrangement of these two components as long as the probe complements the oligonucleotides, and the four outermost nucleotides complement the overhang to be registered at the address of interest.

The target nucleic molecules are distributed over the microarray, together with the fluorescent probe and a nucleic acid ligase with a suitable reaction buffer. When incubating, the target nucleic acid overhang will ligate with the oligonucleotides, provided that the 4 outermost nucleotides are complementary. Thus the fluorescent probes will ligate with the target nucleic acid overhang. By observing if the address fluoresces after washing off unligated probes, one will be able to determine whether the overhang TGCA was present in the target nucleic acid molecule.

In a variation of the above method, multiple overhangs may be registered using the same adaptor. The strategy described above can be extended further to make it possible to register multiple overhangs at the same address. For the above example, one can for instance add the following probes:

1) CGCCTACGTCCT
2) CGCCTACGTCC
3) CGCCTACGTC

The three probes are marked with three different fluorophores - yellow, green, and red, respectively. If the address illuminates yellow when reading, one knows that the probe 1) has been ligated with the 3-nucleotide-long overhang GCA. Accordingly, green fluorescence will indicate that probe 2) has been ligated with the 4-nucleotide-long overhang TGCA, and red fluorescence indicates that probe 3) has been ligated with the 5-nucleotide-long overhang CTGCA.

Thus in some embodiments of the invention, the labelled probes comprise a set of labelled probes, having different lengths and different labels.

In some instances, for example, if one wished to reduce the number of addresses required, it could be useful not to register all bases in an overhang. This may be accomplished by using an adaptor that contains one or more universal bases (U) or by using adaptors with two or more permutations at the same address.

It should be noted that the strategy described above may also be used for sorting. The probe and the oligonucleotide may for example contain a cleavage site for a restriction endonuclease. To ensure that the target DNA molecules are released at the right time in the sorting procedure, the probe has to be attached to the oligonucleotide when the restriction endonuclease performs the cut.

Second Stage - Method 2. A further approach which may be used to identify the nucleic acid fragments which are ligated in the first stage is to use tags which are bound to the second overhang-adaptors.

In this embodiment, the second stage comprises the steps of:

(d1) optionally releasing each population of ligated nucleic acid fragments from the solid support,

selectively contacting each population of nucleic acid fragments which are or were ligated at their first ends to a first overhang-adaptor with a second set of overhang-adaptors,

each overhang-adaptor of the second set comprising a nucleic acid

molecule comprising at least one 5'- or 3'-single-stranded end,

the single-stranded ends of the overhang-adaptors of the second set being of lengths and orientations (i.e. 5'- or 3'-) corresponding to the lengths and orientations of the overhanging single-strands of the cleavage sites of the said restriction endonucleases,

wherein said second set comprises a collection of overhang-adaptors whose single-stranded ends collectively encode up to all possible permutations and combinations of the nucleotides A, C, G and T,

and wherein each different overhang-adaptor in the second set is bound to an individual tag;

(d2) contacting the nucleic acid fragments with a nucleic acid ligase to cause selective ligation of the nucleic acid fragments with those overhang-adaptors of the second set whose 5'- or 3'- single-stranded ends are fully complementary to the second 5'- or 3'-overhanging ends of the nucleic acid fragments;

thus forming a plurality of populations of nucleic acid fragments which are ligated at their second ends to a tagged second overhang-adaptor, and

optionally removing the unligated nucleic acid fragments;

(d3) identifying the sequences of the first and second overhanging ends of each of the nucleic acid fragments from the tags which are bound to the second overhang-adaptors.

The comments given above with regard to the production and use of a second set of overhang-adaptors apply, mutatis mutandis, to this embodiment.

In a preferred aspect of the invention, steps (b)-(d2) are carried out simultaneously, i.e. the nucleic acid fragments are combined with the first and second sets of overhang adaptors simultaneously with the nucleic acid ligase. This provides three different possibilities: 1) Nucleic acid fragments that have been ligated with the first adaptor at one end and the second adaptor at the

other end; 2) Nucleic acid fragments with the first adaptor on both ends; and 3) Nucleic acid fragments with the second adaptors on both ends. It is only the fragments in the first group that will result in successfull signals. Fragments from the other groups will not produce problems, however, since they will either give rise to no signal or be removed during washing.

With regard to this embodiment of the invention, the nucleic acid fragments may be bound or capable of being bound either via the first or the second overhang-adaptors.

The term "tag" is used in this context to refer to a structure or molecule which is capable of representing the sequence information of a pair of first and second overhanging ends of any one nucleic acid fragment; and which is distinguishable from all of the other individual tags.

The tag may be a specific DNA sequence, e.g. 50-500bp long that can be amplified and then used as a probe that is hybridised to a microarray.

Alternatively, the tags may be DNA sequences of different lengths that can be separated, for example on a gel. In this case, the first tags may be amplified (for example by PCR) or released from the solid substrate. It will also normally require that one gel separation is run per well By performing one gel separation per well, this avoids the need for letting the tags represent both overhangs.

Another system for the identification of the tagged second overhang-adaptors is to have tags comprising a group of hybridisation sequences to which a plurality of labelled probes may selectively be hybridised, each group of hybridisation sequences being representative of the sequence of the second overhanging end of the nucleic acid fragment, and each labelled probe being representative of one or more of the nucleotides present in that second overhanging end of the nucleic acid fragment.

Using this system, the group of hybridisation sequences may be read in a number of cycles, in most cases, the number of cycles corresponding essentially to the number of overhanging nucleotides n in the second overhanging end of the nucleic acid fragments.

This stage therefore comprises steps of:

(d1) optionally releasing each population of ligated nucleic acid fragments from the solid support,

selectively contacting each population of nucleic acid fragments which are or were ligated at their first ends to a first overhang-adaptor with a

second set of overhang-adaptors,

each overhang-adaptor of the second set comprising a nucleic acid molecule comprising at least one 5'- or 3'-single-stranded end,

the single-stranded ends of the overhang-adaptors of the second set being of lengths and orientations (i.e. 5'- or 3'-) corresponding to the lengths and orientations of the overhanging single-strands of the cleavage sites of the said restriction endonucleases,

wherein said second set comprises a collection of overhang-adaptors whose single-stranded ends collectively encode up to all possible permutations and combinations of the nucleotides A, C, G and T,

wherein each different overhang-adaptor in the second set is bound to an individual tag;

wherein the tag comprises a plurality of hybridisation sequences, each hybridisation sequence being representative of one or more of the nucleotides in the second overhanging end of the nucleic acid fragment;

(d2) contacting the nucleic acid fragments with a nucleic acid ligase to cause selective ligation of the nucleic acid fragments with those overhang-adaptors of the second set whose 5'- or 3'- single-stranded ends are fully complementary to the second 5'- or 3'-overhanging ends of the nucleic acid fragments;

thus forming a plurality of populations of nucleic acid fragments which are ligated at their second ends to a tagged second overhang-adaptor, and

optionally, removing the unligated nucleic acid fragments;

(d3) contacting the tagged populations of nucleic acid fragments with a set of labelled probes, each set of labelled probes comprising at least one probe which is capable of binding specifically to at least one of the hybridisation sequences;

(d4) identifying which labelled probe has bound to the hybridisation sequence and identifying the spatial position of the bound probe;

(d5) removing the labelled probe from the hybridisation sequence; and

(d6) repeating steps (d3)-(d4), and optionally (d5), until the sequence of the overhang of the second end of the nucleic acid fragment has been determined;

(e) comparing the sequences of the ends of the nucleic acid fragments in order to produce a map of the target nucleic acid molecule.

With regard to this embodiment of the invention, the nucleic acid fragments may be bound or capable of being bound either via the first or the second overhang-adaptors.

The hybridisation sequences may comprise a plurality of discrete or overlapping sequences to which separate labelled probes may be hybridised. Most preferably, the hybridisation sequences are single-stranded DNA sequences.

The labelling of the probes may be by any means which is sufficient to distinguish the labelled probes from each other. Examples of labels include fluorescent moieties.

In most cases, this system will require one labelled probe per overhanging end nucleotide. In some circumstances, however, probes may be used which represent combinations of two or more nucleotides. For example, probes may represent A, C, G or T; or the probes may represent AA, AC, AG, AT, CA, etc.. In the latter case, less cycles will be required, although a larger number (e.g. 16 in this case) of different distinguishable labels will be required.

It is also possible to use a binary system where each nucleotide is represented by two probes.

An example of the above method is given below. In this example, a four nucleotide overhang in the overhang-adaptor may, for example, be read with the tag illustrated below:

Overhang-adaptor Hybridisation sequence

::::::

CGAT:::::: : Sequence 1C : Sequence 2G : Sequence 3A : Sequence 4T

The tag that recognises the overhang GCTA contains a complementary overhang shown to the left and four hybridisation sequences shown to the right. After the overhang-adapter has been ligated to the overhang of the second end of the nucleic acid fragment and attached to the substrate four hybridisation cycles are performed:

First cycle:

:::GGTA:::::: *** Green Probe**

:::CGAT:::::: : Sequence 1C : Sequence 2G : Sequence 3A : Sequence 4T

The labelled green probe binds to the hybridisation sequence that is representative of a C at the first position. Labelled probes which are representative of A, G or T at the first position will not bind.

Second cycle:

:::GGTA:::::: *** Red Probe**

:::CGAT:::::: : Sequence 1C : Sequence 2G : Sequence 3A : Sequence 4T

The labelled red probe binds to the hybridisation sequence that is representative of a G at the second position. Labelled probes which are representative of A, C or T at the second position will not bind.

Third cycle:

:::CGTA:::::: *** Yellow Probe**

:::CGAT:::::: : Sequence 1C : Sequence 2G : Sequence 3A : Sequence 4T

The labelled yellow probe binds to the hybridisation sequence that is representative of an A at the third position. Labelled probes which are representative of C, G or T at the third position will not bind.

Fourth cycle:

:::CGTA:::::: *** Blue Probe**
:::CGAT:::::: : Sequence 1C : Sequence 2G : Sequence 3A : Sequence 4T

The labelled blue probe binds to the hybridisation sequence that is representative of a T at the fourth position. Labelled probes which are representative of A, C or G at the fourth position will not bind.

For each position, there exist four different sequence alternatives representing each of the four nucleotides that can be in the position. The sequence alternatives representing a given nucleotide, such as A, differ between the different positions allowing each position to be analysed independently. In this example, the overhang contains a C in position 1 and hence labelled hybridisation sequence C is used. After adding the four candidate probes that can be hybridised to position 1, a green probe representing C is attached. The probe is washed away after reading and four new candidate probes that can be attached to position 2 is added to the solution, and so on.

The following embodiment provides a method for making maps with multiple restriction endonucleases by performing parallel mapping reactions.

If it is desired to make maps of nucleic acid molecules containing multiple restriction endonuclease sites, it is possible to perform a mapping reaction with enzymes that generate different overhang lengths and qualities (for example EarI, BbvI and BaeI). It will, however, be impossible to distinguish between restriction endonucleases that produce the same kind of overhang, for example BbvI and Alw26I (both of which produce 4-nucleotide 5'-overhangs), if they are used in the same mapping reaction.

The invention therefore provides a method where several mapping reactions are carried out in parallel and, by combining the information from them all, it is possible to generate a consensus map where it is possible to distinguish between restriction endonucleases that produce the same kind of overhang.

The invention therefore provides a method for mapping a nucleic acid molecule comprising the steps of:

(A) treating the nucleic acid molecule with a first set of Type IIs restriction endonucleases to produce one or more nucleic acid fragments, each of the restriction endonucleases in the first set producing different

overhanging single-stranded ends to the other restriction endonucleases in the first set,

and determining the sequences of the overhanging ends of the nucleic acid fragments produced thereby;

(B) treating the nucleic acid molecule with a second set of Type IIs restriction endonucleases to produce one or more nucleic acid fragments, the second set comprising at least one Type IIs restriction endonuclease which was not used in step (A) but which has a cleavage site which is the same as one or more of the Type IIs restriction endonucleases used in step (A);

and determining the sequences of the overhanging ends of the nucleic acid fragments produced thereby;

(C) optionally treating the nucleic acid molecule with one or more further sets of Type IIs restriction endonucleases to produce one or more nucleic acid fragments,

and determining the sequences of the overhanging ends of the nucleic acid fragments produced thereby;

(D) treating the nucleic acid molecule simultaneously with the Type IIs restriction endonucleases from all of the sets to produce one or more nucleic acid fragments,

and determining the sequences of the overhanging ends of the nucleic acid fragments produced thereby;

(E) producing a map of the nucleic acid molecule by using the information derived from steps (A)-(D).

The nucleic acid molecule in this embodiment is preferably a double-stranded DNA molecule. In some embodiments of this invention, step (C) is omitted. In other embodiments, the number of restriction endonucleases used in each set is independently 2, 3 or 4. Thus for example, step (A) may be carried out with 3 restriction endonucleases, step (B) may be carried out

with 3 restriction endonucleases and step (D) may be carried out with 6 restriction endonucleases.

The determining of the sequences of the overhanging ends of the nucleic acid fragments produced in steps (A)-(D) is preferably carried out using a method disclosed herein.

An example of this method is given below:

	5', 3 nucleotide overhang	5', 4 nucleotide overhang	3', 5 nucleotide overhang
Mapping reaction 1	EarI	BbvI	BaeI
Mapping reaction 2	SapI	Alw26I	Bp1I
Mapping reaction 3	EarI + SapI	BbvI + Alw26I	BaeI + Bp1I

Map 1:

EarI	BaeI	EarI	BbvI
(GCT)	(TCTTT)	(TTT)	(GCCT)

Map 2:

Alw26I	Alw26I	SapI	Bp1I
(GTGC)	(TACA)	(TGT)	(AAAGA)

Map 3:

EarI/ SapI	BbvI/ Alw26I	BbvI/ Alw26I	BaeI/ Bp1I	EarI/ SapI	EarI/ SapI	BaeI/ Bp1I	BbvI/ A1w26I
(GCT)	(GTGC)	(TACA)	(TCTTT)	(TTT)	(TGT)	(AAAGA)	(GCCT)

Consensus map:

EarI			BaeI	EarI			BbvI
	Alw26I	Alw26I			SapI	Bp1I	
(GCT)	(GTGC)	(TACA)	(TCTTT)	(TTT)	(TGT)	(AAAGA)	(GCCT)

This strategy can of course be expanded further with additional restriction endonucleases and mapping reactions. It is therefore possible to make detailed maps with 10-20 different Ip and IIs restriction endonucleases. In addition to that, it is also possible to place as much as 10-40 different ordinary Type II enzymes (such as EcoRI and HindIII) into the map as soon as the framework with Type Ip and IIs endonucleases is established.

A further embodiment of the invention provides a method of mapping a target nucleic acid molecule, the method comprising the steps of:

(a) treating the target nucleic acid molecule with one or more restriction endonucleases to produce one or more nucleic acid fragments having first and second 5'- or 3'- single-stranded overhanging ends,

(b) adding the nucleic acid fragments to a first set of overhang-adaptors,

each overhang-adaptor of the first set comprising a nucleic acid molecule comprising at least one 5'- or 3'-single-stranded end,

the single-stranded ends of the first overhang-adaptors being of lengths and orientations (i.e. 5'- or 3'-) corresponding to the lengths and orientations of the overhanging single-strands of the cleavage sites of the said restriction endonucleases,

wherein said first set comprises a collection of overhang-adaptors whose single-stranded ends collectively encode up to all possible permutations and combinations of the nucleotides A, C, G and T at all positions in the single-stranded ends except one or more positions, the latter positions being taken by universal nucleotides,

and wherein each overhang-adaptor in the said first set is spatially separable from every other different overhang-adaptor in the first set;

(c1) contacting the said nucleic acid fragments with a nucleic acid ligase to cause selective ligation of the nucleic acid fragments with those overhang-adaptors whose 5'- or 3'- single-stranded ends are

complementary to the 5'- or 3'-overhanging single-stranded ends of the nucleic acid fragments,

thus forming a plurality of separable populations of nucleic acid fragments which are ligated at their first ends to a first overhang-adaptor, and then

optionally, removing any nucleic acid fragments which are not ligated to first overhang-adaptors;

(c2) releasing the nucleic acid fragments which are bound at their first ends with a restriction endonuclease which creates a new first overhanging single-stranded end in the nucleic acid fragment which comprises the nucleotide or nucleotides in the nucleic acid fragments which corresponded to the universal nucleotides;

(d1) selectively contacting each released population of nucleic acid fragments with a second set of overhang-adaptors,

each overhang-adaptor of the second set comprising a nucleic acid molecule comprising at least one 5'- or 3'-single-stranded end,

the single-stranded ends of the overhang-adaptors of the second set being of lengths and orientations (i.e. 5'- or 3'-) corresponding to the lengths and orientations of the overhanging single-strands of the cleavage sites of the said restriction endonuclease,

wherein said second set comprises a collection of overhang-adaptors whose single-stranded ends collectively encode up to all possible permutations and combinations of the nucleotides A, C, G and T,

and wherein each overhang-adaptor in the said second set is spatially distinguishable from every other different overhang-adaptor in the second set;

(d2) contacting the nucleic acid fragments with a nucleic acid ligase to cause selective ligation of the nucleic acid fragments with those

overhang-adaptors of the second set whose 5'- or 3'- single-stranded ends are fully complementary to the second 5'- or 3'-overhanging ends of the nucleic acid fragments;

thus forming a plurality of populations of nucleic acid fragments which are ligated at their second ends to a second overhang-adaptor, and then

optionally, removing any unbound nucleic acid fragments;

(d3) contacting the ligated nucleic acid fragments with labelled-adaptors which bind selectively to the new first overhanging end on the basis of the nucleotide or nucleotides in the new first overhanging end of the nucleic acid fragments which corresponded to the universal nucleotides;

(d4) identifying the sequences of the first and second overhanging ends of each of the nucleic acid fragments from the spatial positions of the second overhang-adaptors to which the nucleic acid fragments are ligated, and from the labels which are attached to the first ends of the nucleic acid fragments; and

(e) comparing the sequences of the ends of the nucleic acid fragments in order to produce a map of the target nucleic acid molecule.

This method has the advantage that an initial sorting is carried out using a smaller number of first overhang-adaptors. The missing sequence information is retrieved by making us of labelled-adaptors in the second stage. This embodiment is illustrated in Figures 2 and 3.

In this embodiment, the first set comprises a collection of overhang-adaptors whose single-stranded ends collectively encode up to all possible permutations and combinations of the nucleotides A, C, G and T at all positions in the single-stranded ends except one or more positions, the latter positions being taken by universal nucleotides. In this context, "universal nucleotides" are nucleotides which are capable of base-pairing with any of the nucleotides A, C, G and T.

Preferably, universal nucleotides are present at one or two positions in the single-stranded ends of the first overhang-adaptors.

The ligated nucleic acid fragments are released from the first overhang-adaptors by cleavage with a restriction endonuclease which creates a new first overhanging single-stranded end in the nucleic acid fragment. It should be noted that the new first overhanging end in the nucleic acid fragment may be the same length and orientation as the initial first overhanging end. Preferably this is done with a Type Ip or Type IIs restriction endonunclease, for example, one having a recognition site in the first overhang-adaptor. The new first overhanging end comprises the nucleotide or nucleotides in the nucleic acid fragments which corresponded to the universal nucleotides, i.e. those nucleotides which hybridised opposite or base-paired with the universal nucleotides.

The comments given above with regard to overhang-adaptors of the first and second sets apply herein, mutatis mutandis.

The binding of the second overhanging ends of the nucleic acid fragments to the second overhang-adaptors will provide the majority of the sequence information on the first overhanging end of the nucleic acid fragment and all of the sequence information on the second overhanging end. The remaining information on the sequence of the first overhanging end is obtained through the use of labelled-adaptors.

Labelled-adaptors are used which bind selectively to the new first overhanging ends of the nucleic acid fragments on the basis of the nucleotide or nucleotides in the new first overhanging ends of the nucleic acid fragments which corresponded to the universal nucleotides in the first overhang-adaptors. Thus the labelled adaptors will provide the information on the sequence of the first overhanging ends which was not provided by the binding of the nucleic acid fragment to the first overhang-adaptor.

Preferably, the set of labelled-adaptors comprises a set of four labelled adaptors, wherein the labels are distinguishable from one another. Preferably,the labels are fluorescent moieties.

In the context of this invention, labels are those which directly or indirectly allow detection and/or determination by the generation of a signal. Such labels include for example radiolabels, chemical labels (e.g. EtBr, TOTO, YOYO and other dyes), chromophores or fluorophores (e.g. dyes such as fluorescein and rhodamine), or reagents of high electron density such as ferritin, haemocyanin or colloidal gold. Alternatively, the label may be an enzyme, for example peroxidase or alkaline phosphatase, wherein the presence of the enzyme is visualized by its interaction with a suitable entity, for example a substrate. The label may also form part of a signalling pair

wherein the other member of the pair may be introduced into close proximity, for example, a fluorescent compound and a quench fluorescent substrate may be used.

A label may also be provided on a different entity, such as an antibody, which specifically recognizes at least a region of molecule to be identified. If the molecule to be identified is a polynucleotide, one way in which a label may be introduced for example is to bind a suitable binding partner carrying a label, e.g. fluorescent labelled probes or DNA-binding proteins.

Once the sequence information has been accumulated, a computer program may then be used to assemble the sequence pieces into the final sequence.

Kits for performing the mapping methods described herein form a further aspect of the invention. Thus viewed from a further aspect, the present invention provides a kit for mapping a target nucleic acid molecule comprising a set of first overhang-adaptors as described herein, optionally attached to one or more solid supports; a set of second overhang-adaptors as described herein; and one or more restriction endonucleases for use with one or more of the methods described herein.

Optionally the kit may contain other appropriate components selected from the list including vectors into which the target molecules may be ligated, ligases, enzymes necessary for inactivation and activation of restriction or ligation sites, primers for amplification and/or appropriate enzymes, buffers and solutions. Appropriate labelling means may also be included in such kits.

The use of such kits for mapping target nucleic acid molecules form further aspects of the invention.

To increase the statistical capacity of the method, it is possible to record restriction endonuclease cleavage sites located between the two overhanging ends of a nucleic acid fragment. One strategy is to free DNA molecules from the wells used in the first sorting step using restriction endonucleases which do not have binding sites in the overhang-adaptors. There must be cleavage sites in the actual target DNA for the DNA molecules to be freed and made available for the next step. This procedure may of course be repeated with several restriction endonucleases. It is also possible to cut with several endonucleases at once. If labelled adaptors are then used that recognise and label the different overhangs with different colours, it will be possible to record which enzyme has freed the nucleic acid fragment.

Thus the invention relates to a method wherein after the nucleic

acid fragments are selectively ligated to the first overhang-adaptors, they are treated with a restriction endonuclease; and a labelled adaptor is then used to determine whether the restriction endonuclease has cut the nucleic acid fragment. The labelled adaptor may bind either to the cut end of the released nucleic acid fragment or to the cut end of the bound nucleic acid fragment.

The presence or absence of a restriction endonuclease cleavage site in a target nucleic acid molecule may also be determined by immobilising the target nucleic acid molecule on a solid support, for example, a microarray, and to label the free end of the target nucleic acid, for example with a fluorescent moiety. The target nucleic acid molecule may then be treated with a restriction endonuclease. If the label disappears after the restriction endonuclease treatment, then it can be said that the restriction endonuclease cuts in the target nucleic acid. This method is illustrated in Figure 4.

The invention therefore provides a method for determining the presence or absence of a restriction endonuclease cleavage site within a target nucleic acid comprising the steps of immobilising the target nucleic acid molecule on a solid support, labelling the free end of the target nucleic acid; treating the target nucleic acid with a restriction endonuclease; and then determining the presence or absence of the label after treatment.

This procedure may be repeated with several restriction endonucleases. Similarly, it is possible to cut with several restriction endonucleases at once and label the different overhangs with different colours.

It is also possible to extend the last-mentioned principle to sequencing. In this method, one end of a target nucleic acid molecule is ligated with a linker containing a Type Ip or Type IIs restriction endonuclease recognition site which creates an overhang in the actual target nucleic acid molecule of one or more bases. The overhang in the target nucleic acid molecule is then ligated with a labelled adaptor that recognises one or more overhanging bases. It is possible, for example, to use four different labelled adaptors that recognise adenine, cytosine, guanine and thymine, and which are labelled with four different fluorescent colours. The fluorescent colour of the address thus provides information on which base is in the position being analysed.

If the labelled adaptors contain a cleavage site for a Type Ip or Type IIs restriction endonuclease that generates a new overhang in the target sequence, and which has been displaced in relation to the first overhang, the process can be repeated one or more times, providing sequence information

in towards the centre of the target nucleic acid sequence in a controlled manner.

There is provided therefore a method of sequencing a target nucleic acid molecule comprising the steps of:

(i) ligating the target nucleic acid molecule with a linker nucleic acid, the linker nucleic acid comprising a recognition site for a Type Ip or Type IIs restriction endonuclease which will cleave the target nucleic acid molecule;

(ii) treating the target nucleic acid molecule with a Type Ip or Type IIs restriction endonuclease to produce one or more nucleic acid fragments having single-stranded overhanging ends;

(iii) ligating one or more of the target nucleic acid fragments with a set of labelled adaptors which specifically recognise one or more of the nucleotides in the single-stranded overhanging ends of the nucleic acid fragments, wherein the labelled adaptors comprise a recognition site for a Type Ip or Type IIs restriction endonuclease which will cleave the target nucleic acid molecule at a position one or more nucleotides 5'- or 3'- to the first cleavage site;

(iv) identifying which labelled adaptors have bound to the nucleic acid fragments, thus providing information on the nucleotide sequence of at least part of the overhanging ends of the target nucleic acid fragment;

(v) optionally, repeating steps (ii)-(iv) one or more times.

Preferably, the nucleic acid molecule is a DNA molecule, most preferably a double-stranded DNA molecule.

In a particularly preferred embodiment, the target DNA is immobilised at one end, for example on an array, and the linkers and labelled adaptors are bound to the other free end.

The above strategy may provide several important benefits over methods known in the prior art. Firstly, the use of microarrays means that the analyses are not based on signals from single molecules, but from a large set of equal-length target sequences. Stronger signals may therefore be obtained compared with scanning strategies that are based on single molecules. Furthermore, it is easier to carry out a lot of cycles as loss of target DNA can be tolerated.

It should be noted that in all embodiments of the invention, the nucleic acid fragments may be amplified by a linear or exponential PCR using

the first and/or second overhang-adpators as PCR primers.

LEGENDS TO FIGURES

Figure 1 Method for registering overhangs on a microarray address.

Figure 2 Example of the use of multiple coloured fluorescence adaptors in order to reduce the number of addresses required.

Figure 3 Example of the identification of internal cleavage sites, illustrating the presence of the doublet AAAT-CAGA.

Figure 4 Example of how the fluorescent colour disappears from an address if the DNA fragment contains a cleavage site for the restriction endonucleases being used.

Figure 5 Digestion of a target nucleic acid molecule with Type IIs restriction endonucleases Fok I and Hga I.

Figure 6 Example of the first stage of the mapping procedure using a restriction endonuclease which produces a 4 nucleotide single-stranded overhang.

Figure 7 Example of an area of from a microarray, illustrating the presence of the doublet TTTA-GTCT

The following examples are given by way of illustration only and should not be read as limiting the invention in any way.

EXAMPLES

Example 1 - Mapping method

The mapping principle is illustrated by the Type IIs restriction endonucleases HgaI and FokI as shown in Figure 1.

In the first step, the target sequence was cut with HgaI and FokI to form five fragments each with two unique overhanging ends. This included a FokI site with an ACGT overhang furthest to the left. This was followed by an HgaI site, three FokI sites and, finally, an HgaI site furthest to the right. The map produced contained the internal sequence of the two restriction endonucleases, together with details of the sequences and positions of the overhanging ends.

Example 2 - Scanning using microarrays

The mapping procedure was carried out with FokI, i.e. an enzyme that creates 4-nucleotide overhangs. With such an enzyme, 256 overhang permutations can be generated, which in turn means that there are 256 x 256 permutations of overhanging end pairs. A microarray with 65,536 addresses was thus used to identity the overhanging end pairs present in a solution.

Several strategies were envisaged for assigning the overhang pairs to the correct addresses in the microarray. In this case, ligations and a two-step sorting procedure as illustrated in Figure 6 were used.

We started with a microtitre plate with 256 wells including overhang adaptors anchored to the wells' substrates. Well 1 contained adapters with AAAA overhangs, well 2 contained adapters with AAAC overhangs etc., so that each overhang permutation had its own well. The solution with the overhang pairs was then distributed evenly between the wells. Ligase was added so that the pairs with overhangs complementing the overhang-adapters in the respective wells were ligated (the overhang pairs were treated with phosphatase initially in order to reduce the occurrence of ligations between overhang pairs). Then, after washing the well, we were left with just two overhang pairs which had been ligated. These were then freed, by means of a cleavage site located in the overhang adaptors, so we could then proceed to the next sorting round.

It should be noted that the overhangs that were ligated to the overhang adaptors had now been inactivated. Freed DNA molecules from well 1 were then added to area 1 on a microarray, DNA molecules from well 2 to area 2, etc.. The 256 areas on the microarray were physically separated from each other. Furthermore, each area was divided into 256 addresses, address no. 1 comprising overhang adaptors with AAAA overhangs, address no. 2 has AAAC overhangs etc.. We then incubated the DNA solution with ligase, and the overhang pairs with TTTT overhangs ligated to address 1, and so on.

After the overhang pairs were ligated to their respective addresses, the microarray could be scanned. The information was scanned as shown in Figure 7.

We recorded a light signal at address 85, area 4, hence we knew that one overhang must be TTTA because all the DNA in this area was sorted into well no. 4 where overhang adaptors with AAAT overhangs were used. Similarly we could ascertain that the other overhang must have been

GTCT as the overhang adaptors at address 85 have CAGA overhangs.

Example 3 - Mapping of pBluescript

In this example, a combination of Type IIs and Ip restriction endonucleases were used to map DNA sequences.

In this procedure, the mapping of HgaI, FokI and BstXI on pBluescript is used as an example. In the procedure, pBluescript is digested with the three enzymes which generates 9 fragments. A complete set of adapters is ligated to the overhangs (in this example, they are called left and right adapters for convenience). The left adapters will recognize the left overhangs and the right adapters, the right overhangs. Ligation is performed in 9 tubes where each tube contains a specific biotinylated left adapter corresponding to a specific overhang. By adding streptavidin-coated beads to the wells a sorting based on the left overhangs is performed. After extensive washing to remove unbound fragments, a linear PCR (or alternatively, exponential PCR) is performed on the right adapter which is ligated on to the other overhang. It should be noted that the sequence of the right adapter, except for a common primer site, is specific for the right overhang it recognizes. The adapter, and thus, the overhang sequence, can be determined by hybridization to its counterpart on a microarray. Based on the overhang quality, the order of restriction endonucleases can be mapped on pBluescript.

<u>Reagents:</u>

pBluescript SKII+
HgaI (NEB)
FokI (NEB)
BstXI (NEB)
Biotinylated left adapters (~30 bp)
Non-biotinylated left adapters (~30 bp)
Right adapters (~90 bp)
T4 DNA ligase buffer (NEB)
T4 DNA ligase (NEB)
Taq polymerase buffer (Dynazyme)
dNTPs
Taq (Dynazyme)
Polylysine-coated slides
Cy3-labelled antisense right adapter oligo

Hybridization solution
Streptavidin-coated M-270 beads
Bind and wash buffer (B&W) (10 mM Tris-Cl, pH 7.5, 1 mM EDTA, 2M NaCl)

Protocol:

Step I: Digestion of pBluescript:

First digestion:

pBluescript:	18 mg
NEB3 buffer:	1X
BstXI:	36U
Volume:	50 ml

Incubation at 55°C for 1 hr. Ethanol precipitation to change buffer.

Second digestion:

BstXI digested pBluescript:	18 mg
NEB4:	1X
HgaI:	36U
FokI:	36U
Volume:	270 ml

Incubation at 37°C for 1 hr. Ethanol precipitation to concentrate the sample. Dissolvation of sample to 1 mg/ml by adding 18 ml of TE. This concentration equals 0.52 pmol/ml of pBluescript.

Step II: Ligation of adapters to pBluescript fragments:

9 tubes containing the following:

Tubei (where i=A-I)	
Digested pBluescript (from step I):	1 pmol (= 2 mg)
Ligase buffer:	1X
All left adapters - adapteri:	10 pmol each

Biotinylated adapteri:	10 pmol
All right adapters:	10 pmol each
T4 DNA ligase:	800 U
Volume:	60 ml

Incubation at 20°C for 4 hrs.

Step III: Immobilization

Mix each of the tubes from step II with:

Equilibrated M-270 beads:	0.1 mg
Volume (2X B&W):	60 ml

Incubation at 25°C for 1 hr with rotation (rotator).

Three washes using 120 ml 2X B&W buffer. Additional wash with 120 ml 1X PCR buffer. Beads dissolved in 10 ml 1X PCR buffer.

Step IV: PCR amplification of right adapter using Cy3-primer

Taq buffer:	0.8X
$MgCl_2$:	6 mM
dNTPs:	50 mM
Cy3-primer:	100 pmol
Template on beads in 1X buffer: (1 pmol immobilized fragment)	10 ml
Taq polymerase:	0.4U
Volume:	50 ml

Thermal cycling: 95°C, 2 min; 95°C, 15 sec, 58°C, 30 sec, 72°C, 15 sec; 30 cycles

Ethanol precipitation of PCR product to increase concentration.

Step V: Hybridization of PCR-amplified probes to microarrays

Each PCR-amplified probe is hybridized to a separate microarray. Each

microarray printed on poly-L-lysine coated slides in exact same way and includes 6 control spots (Cy3-labelled oligo) and 9 test spots. Each of the 9 test spots contains oligo that is supposed to hybridize only to the PCR product from a single adaptor template (the list of oligos)

Hybridization is to be carried out according to the following protocol:

1. Each PCR-amplified probe (containing Cy3-label) is to be dissolved in 1,7 ml of 2X hybridization solution (7X SSC and 0.6% SDS) and added the following mix to get hybridization probes:

50x Denhardt's reagent:	0.5 ml
tRNA (4mg/ml):	0.5 ml
Salmon sperm DNA (10mg/ml):	0.5 ml
2x hybridization solution:	3.3 ml
Water:	3.5 ml
Total hybridization volume	10.0 ml

2. Poly-L-lysine coated slides with printed microarrays are to be pre-processed according to protocol from P. O. Brown's laboratory:

a. DNA is to be cross-linked to the slides by irradiating with UV (60 mJ)
b. Slides are to be blocked in blocking solution (blocking solution contains 6 gram of succinic anhydride dissolved in 335 ml of 1-methyl-2 pyrrolidinone and supplemented with 15 ml of boric acid, pH=8.0) for 20 minutes with vigorous agitation, rinsed in distilled water, boiled in distilled water for 2 minutes, and washed for 2 minutes in cold 96% ethanol.
c. Right before the hybridization, hybridization probes are to be boiled for 2-5 minutes
d. Hybridization probes are to be applied to individual microarrays and hybridized for 12-16 hours under cover slip in humidified chamber(s) inside hybridization oven or in a water bath at 50 - 65°C

3. After hybridization, slides with microarrays are to be removed from the humidified chamber(s) and washed as follows:

a. Once with 1xSSC, 0.05% SDS for 2-3 min
b. Once with 0.2xSSC for 2-3 min

c. Once with 0.05xSSC for 2-3 minutes

4. Slides are then to be dried by gentle centrifugation (1,000 rpm, 5 min)

5. Slides are then to be scanned with a laser appropriate for Cy3 label.

CLAIMS

1. A method of mapping a target nucleic acid molecule, the method comprising the steps of:

(a) treating the target nucleic acid molecule with one or more restriction endonucleases to produce one or more nucleic acid fragments having first and second 5'- or 3'- single-stranded overhanging ends,

(b) adding the nucleic acid fragments to a first set of overhang-adaptors,

each overhang-adaptor of the first set comprising a nucleic acid molecule comprising at least one 5'- or 3'-single-stranded end,

the single-stranded ends of the overhang-adaptors being of lengths and orientations (i.e. 5'- or 3'-) corresponding to the lengths and orientations of the overhanging single-strands of the cleavage sites of the said restriction endonucleases,

wherein said first set comprises a collection of overhang-adaptors whose single-stranded ends collectively encode up to all possible permutations and combinations of the nucleotides A, C, G and T,

and wherein each overhang-adaptor in the said first set is spatially separable from every other different overhang-adaptor in the first set;

(c) contacting the said nucleic acid fragments with a nucleic acid ligase to cause selective ligation of the nucleic acid fragments with those overhang-adaptors whose 5'- or 3'- single-stranded ends are fully complementary to the 5'- or 3'-overhanging single-stranded ends of the nucleic acid fragments,

thus forming a plurality of separable populations of nucleic acid fragments which are ligated at their first ends to a first overhang-adaptor;

optionally, removing the unligated nucleic acid fragments;

(d) identifying the sequence of the second overhanging single-stranded end of the nucleic acid fragments; and

(e) comparing the sequences of the ends of the nucleic acid fragments in order to produce a map of the target nucleic acid molecule.

2. A method as claimed in claim 1 wherein the target nucleic acid molecule is a DNA molecule.

3. A method as claimed in claim 1 or claim 2 wherein the restriction endonuclease is a Type Ip or Type IIs restriction endonuclease.

4. A method as claimed in any one of the previous claims wherein the target nucleic acid molecule is treated with more than one restriction endonuclease, wherein the restriction endonucleases either all produce 5'-overhanging ends or all produce 3'-overhanging ends.

5. A method as claimed in any one of the previous claims wherein the overhang adaptor of the first set are attached or capable of being attached to a solid support.

6. A method as claimed in any one of the previous claims wherein the ligation reaction in step (c) is carried out in free solution.

7. A method as claimed in any one of the previous claims wherein step (d) is carried out by:

(d1) optionally releasing each population of ligated nucleic acid fragments from the solid support,

selectively contacting each population of nucleic acid fragments which were ligated at their first ends to a first overhang-adaptor with a second set of overhang-adaptors,

each overhang-adaptor of the second set comprising a nucleic acid molecule comprising at least one 5'- or 3'-single-stranded end,

the single-stranded ends of the overhang-adaptors of the second set being of lengths and orientations (i.e. 5'- or 3'-) corresponding to the lengths and orientations of the overhanging single-strands of the cleavage sites of the said restriction endonucleases,

wherein said second set comprises a collection of overhang-adaptors whose single-stranded ends collectively encode up to all possible permutations and combinations of the nucleotides A, C, G and T,

and wherein each overhang-adaptor in the said second set is spatially distinguishable from every other different overhang-adaptor in the second set;

(d2) contacting the nucleic acid fragments with a nucleic acid ligase to cause selective ligation of the nucleic acid fragments with those overhang-adaptors of the second set whose 5'- or 3'- single-stranded ends are fully complementary to the second 5'- or 3'-overhanging ends of the nucleic acid fragments;

thus forming a plurality of populations of nucleic acid fragments which are ligated at their second ends to a second overhang-adaptor, and

optionally removing the non-ligated nucleic acid fragments;

(d3) identifying the sequences of the first and second overhanging ends of each of the nucleic acid fragments from the spatial positions of the second overhang-adaptors to which the nucleic acid fragments are ligated.

8. A method as claimed in claim 7 wherein steps (b)-(d2) are carried out essentially simultaneously.

9. A method for detecting overhangs on a microarray address, the method comprising the steps of:

providing one or more single-stranded nucleic acid adaptors each

comprising a first part and a second part, the first and second parts being contiguous with one another, the first part having a free 5'- or 3'-end;

wherein the adaptor is preferably bound to a solid support;

contacting the adaptor with a target nucleic acid molecule having a single-stranded overhang which is complementary with the first part of the adaptor;

ligating the first part of the adaptor to the single-stranded overhang of the target nucleic acid molecule;

contacting the second part of the adaptor with one or more labelled single-stranded nucleic acid probes having a nucleotide sequence which is complementary with the second part of the adaptor;

ligating the labelled single-stranded nucleic acid probe to the target nucleic acid molecule;

optionally removing any unligated labelled single-stranded nucleic acid probe and/or unligated nucleic acid molecule;

determining whether any target nucleic acid molecule has been ligated to the first part of the adaptor by determining whether any labelled probe is bound to the second part of the adaptor.

10. A method as claimed in claim 7, wherein the spatial positions of the second overhang adaptors are determined using the method claimed in claim 8.

11. A method as claimed in any one of claims 1 to 6, wherein step (d) is carried out by:

(d1) optionally releasing each population of ligated nucleic acid fragments from the solid support,

selectively contacting each population of nucleic acid fragments

which are or were ligated at their first ends to a first overhang-adaptor with a second set of overhang-adaptors,

each overhang-adaptor of the second set comprising a nucleic acid molecule comprising at least one 5'- or 3'-single-stranded end,

the single-stranded ends of the overhang-adaptors of the second set being of lengths and orientations (i.e. 5'- or 3'-) corresponding to the lengths and orientations of the overhanging single-strands of the cleavage sites of the said restriction endonucleases,

wherein said second set comprises a collection of overhang-adaptors whose single-stranded ends collectively encode up to all possible permutations and combinations of the nucleotides A, C, G and T,

and wherein each different overhang-adaptor in the second set is bound to an individual tag;

(d2) contacting the nucleic acid fragments with a nucleic acid ligase to cause selective ligation of the nucleic acid fragments with those overhang-adaptors of the second set whose 5'- or 3'- single-stranded ends are fully complementary to the second 5'- or 3'-overhanging ends of the nucleic acid fragments;

thus forming a plurality of populations of nucleic acid fragments which are ligated at their second ends to a tagged second overhang-adaptor, and

optionally removing the unligated nucleic acid fragments;

(d3) identifying the sequences of the first and second overhanging ends of each of the nucleic acid fragments from the tags which are bound to the second overhang-adaptors.

12. A method as claimed in claim 11, wherein the tag is a DNA molecule

13. A method as claimed in any one of claims 1 to 6, wherein step (d)

comprises:

(d1) optionally releasing each population of ligated nucleic acid fragments from the solid support,

selectively contacting each population of nucleic acid fragments which are or were ligated at their first ends to a first overhang-adaptor with a second set of overhang-adaptors,

each overhang-adaptor of the second set comprising a nucleic acid molecule comprising at least one 5'- or 3'-single-stranded end,

the single-stranded ends of the overhang-adaptors of the second set being of lengths and orientations (i.e. 5'- or 3'-) corresponding to the lengths and orientations of the overhanging single-strands of the cleavage sites of the said restriction endonucleases,

wherein said second set comprises a collection of overhang-adaptors whose single-stranded ends collectively encode up to all possible permutations and combinations of the nucleotides A, C, G and T,

wherein each different overhang-adaptor in the second set is bound to an individual tag;

wherein the tag comprises a plurality of hybridisation sequences, each hybridisation sequence being representative of one or more of the nucleotides in the second overhanging end of the nucleic acid fragment;

(d2) contacting the nucleic acid fragments with a nucleic acid ligase to cause selective ligation of the nucleic acid fragments with those overhang-adaptors of the second set whose 5'- or 3'- single-stranded ends are fully complementary to the second 5'- or 3'-overhanging ends of the nucleic acid fragments;

thus forming a plurality of populations of nucleic acid fragments which are ligated at their second ends to a tagged second overhang-

adaptor, and

optionally, removing the unligated nucleic acid fragments;

(d3) contacting the tagged populations of nucleic acid fragments with a set of labelled probes, each set of labelled probes comprising at least one probe which is capable of binding specifically to at least one of the hybridisation sequences;

(d4) identifying which labelled probe has bound to the hybridisation sequence and identifying the spatial position of the bound probe;

(d5) removing the labelled probe from the hybridisation sequence; and

(d6) repeating steps (d3)-(d4), and optionally (d5), until the sequence of the overhang of the second end of the nucleic acid fragment has been determined.

14. A method for mapping a nucleic acid molecule comprising the steps of:

(A) treating the nucleic acid molecule with a first set of Type IIs restriction endonucleases to produce one or more nucleic acid fragments, each of the restriction endonucleases in the first set producing different overhanging single-stranded ends to the other restriction endonucleases in the first set,

and determining the sequences of the overhanging ends of the nucleic acid fragments produced thereby;

(B) treating the nucleic acid molecule with a second set of Type IIs restriction endonucleases to produce one or more nucleic acid fragments, the second set comprising at least one Type IIs restriction endonuclease which was not used in step (A) but which has a cleavage site which is the same as one or more of the Type IIs restriction endonucleases used in step (A);

and determining the sequences of the overhanging ends of the nucleic acid fragments produced thereby;

(C) optionally treating the nucleic acid molecule with one or more further sets of Type IIs restriction endonucleases to produce one or more nucleic acid fragments,

and determining the sequences of the overhanging ends of the nucleic acid fragments produced thereby;

(D) treating the nucleic acid molecule simultaneously with the Type IIs restriction endonucleases from all of the sets to produce one or more nucleic acid fragments,

and determining the sequences of the overhanging ends of the nucleic acid fragments produced thereby;

(E) producing a map of the nucleic acid molecule by using the information derived from steps (A)-(D).

15. A method as claimed in claim 14 wherein the nucleic acid molecule is a DNA molecule.

16. A method of mapping a target nucleic acid molecule, the method comprising the steps of:

(a) treating the target nucleic acid molecule with one or more restriction endonucleases to produce one or more nucleic acid fragments having first and second 5'- or 3'- single-stranded overhanging ends,

(b) adding the nucleic acid fragments to a first set of overhang-adaptors,

each overhang-adaptor of the first set comprising a nucleic acid molecule comprising at least one 5'- or 3'-single-stranded end,

the single-stranded ends of the first overhang-adaptors being of lengths and orientations (i.e. 5'- or 3'-) corresponding to the lengths and orientations of the overhanging single-strands of the cleavage sites of the said restriction endonucleases,

wherein said first set comprises a collection of overhang-adaptors whose single-stranded ends collectively encode up to all possible permutations and combinations of the nucleotides A, C, G and T at all positions in the single-stranded ends except one or more positions, the latter positions being taken by universal nucleotides,

and wherein each overhang-adaptor in the said first set is spatially separable from every other different overhang-adaptor in the first set;

(c1) contacting the said nucleic acid fragments with a nucleic acid ligase to cause selective ligation of the nucleic acid fragments with those overhang-adaptors whose 5'- or 3'- single-stranded ends are complementary to the 5'- or 3'-overhanging single-stranded ends of the nucleic acid fragments,

thus forming a plurality of separable populations of nucleic acid fragments which are ligated at their first ends to a first overhang-adaptor, and then

optionally, removing any nucleic acid fragments which are not ligated to first overhang-adaptors;

(c2) releasing the nucleic acid fragments which are bound at their first ends with a restriction endonuclease which creates a new first overhanging single-stranded end in the nucleic acid fragment which comprises the nucleotide or nucleotides in the nucleic acid fragments which corresponded to the universal nucleotides;

(d1) selectively contacting each released population of nucleic acid fragments with a second set of overhang-adaptors,

each overhang-adaptor of the second set comprising a nucleic acid molecule comprising at least one 5'- or 3'-single-stranded end,

the single-stranded ends of the overhang-adaptors of the second set being of lengths and orientations (i.e. 5'- or 3'-) corresponding to the lengths and orientations of the overhanging single-strands of the

cleavage sites of the said restriction endonuclease,

wherein said second set comprises a collection of overhang-adaptors whose single-stranded ends collectively encode up to all possible permutations and combinations of the nucleotides A, C, G and T,

and wherein each overhang-adaptor in the said second set is spatially distinguishable from every other different overhang-adaptor in the second set;

(d2) contacting the nucleic acid fragments with a nucleic acid ligase to cause selective ligation of the nucleic acid fragments with those overhang-adaptors of the second set whose 5'- or 3'- single-stranded ends are fully complementary to the second 5'- or 3'-overhanging ends of the nucleic acid fragments;

thus forming a plurality of populations of nucleic acid fragments which are ligated at their second ends to a second overhang-adaptor, and then

optionally, removing any unbound nucleic acid fragments;

(d3) contacting the ligated nucleic acid fragments with labelled-adaptors which bind selectively to the new first overhanging end on the basis of the nucleotide or nucleotides in the new first overhanging end of the nucleic acid fragments which corresponded to the universal nucleotides;

(d4) identifying the sequences of the first and second overhanging ends of each of the nucleic acid fragments from the spatial positions of the second overhang-adaptors to which the nucleic acid fragments are ligated, and from the labels which are attached to the first ends of the nucleic acid fragments; and

(e) comparing the sequences of the ends of the nucleic acid fragments in order to produce a map of the target nucleic acid molecule.

17. A method as claimed in claim 16, wherein universal nucleotides are present at one or two positions in the single-stranded ends of the first overhang-adaptors.

18. A method as claimed in claim 16, wherein in step (a), the target nucleic acid molecule is treated with one or more Type Ip or IIs restriction endonucleases.

19. A method of sequencing a target nucleic acid molecule comprising the steps of:

(i) ligating the target nucleic acid molecule with a linker nucleic acid, the linker nucleic acid comprising a recognition site for a Type Ip or Type IIs restriction endonuclease which will cleave the target nucleic acid molecule;

(ii) treating the target nucleic acid molecule with a Type Ip or Type IIs restriction endonuclease to produce one or more nucleic acid fragments having single-stranded overhanging ends;

(iii) ligating one or more of the target nucleic acid fragments with a set of labelled adaptors which specifically recognise one or more of the nucleotides in the single-stranded overhanging ends of the nucleic acid fragments, wherein the labelled adaptors comprise a recognition site for a Type Ip or Type IIs restriction endonuclease which will cleave the target nucleic acid molecule at a position one or more nucleotides 5'- or 3'- to the first cleavage site;

(iv) identifying which labelled adaptors have bound to the nucleic acid fragments, thus providing information on the nucleotide sequence of at least part of the overhanging ends of the target nucleic acid fragment;

(v) optionally, repeating steps (ii)-(iv) one or more times.

20. A method as claimed in claim 19, wherein the target nucleic acid molecule is a DNA molecule.

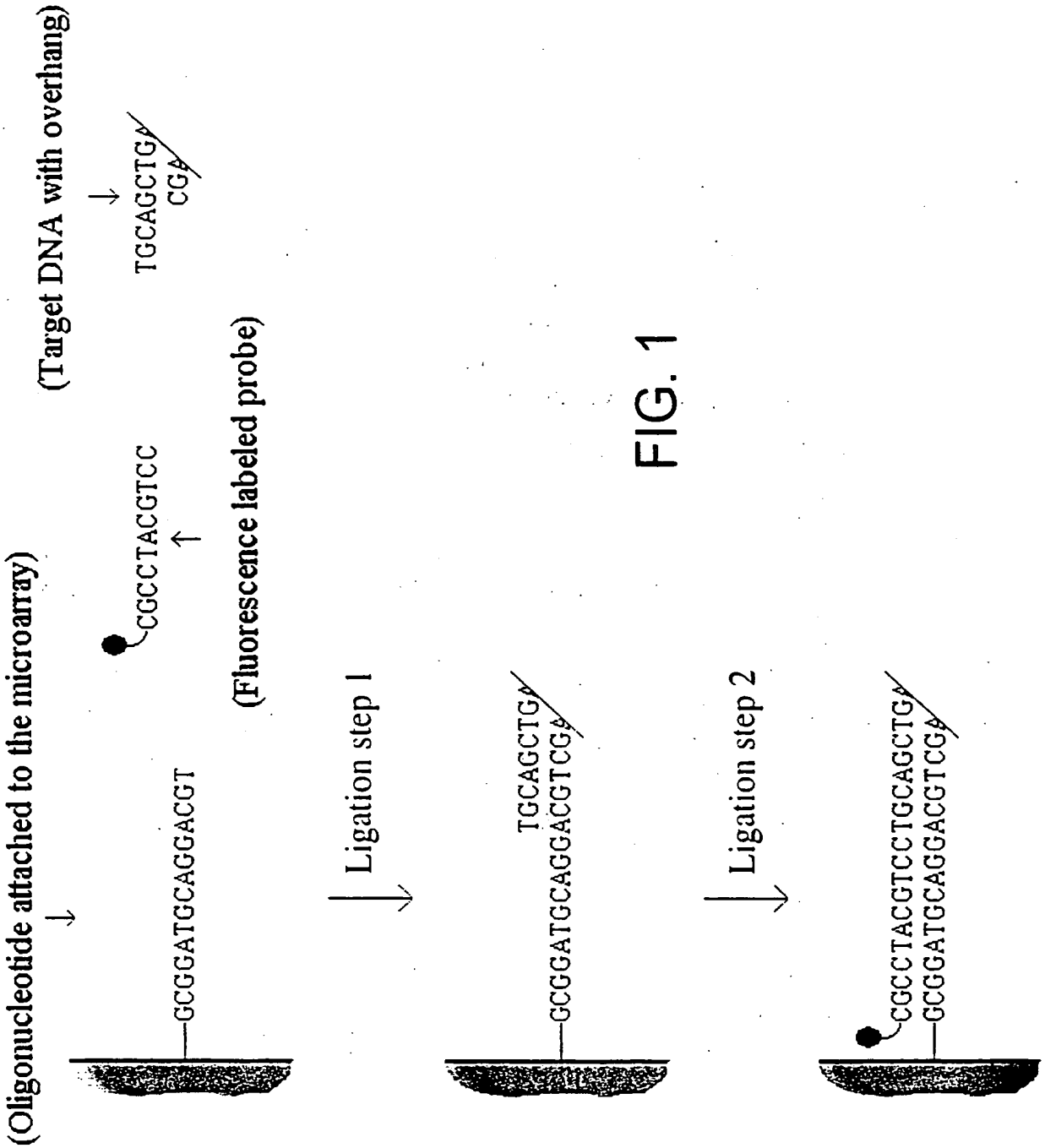

(Oligonucleotide attached to the microarray)
(Target DNA with overhang)
TGCAGCTG
CGA
CGCCTACGTCC
(Fluorescence labeled probe)
GCGGATGCAGGACGT
Ligation step 1
TGCAGCTG
GCGGATGCAGGACGTCGA
Ligation step 2
CGCCTACGTCCTGCAGCTG
GCGGATGCAGGACGTCGA

FIG. 1

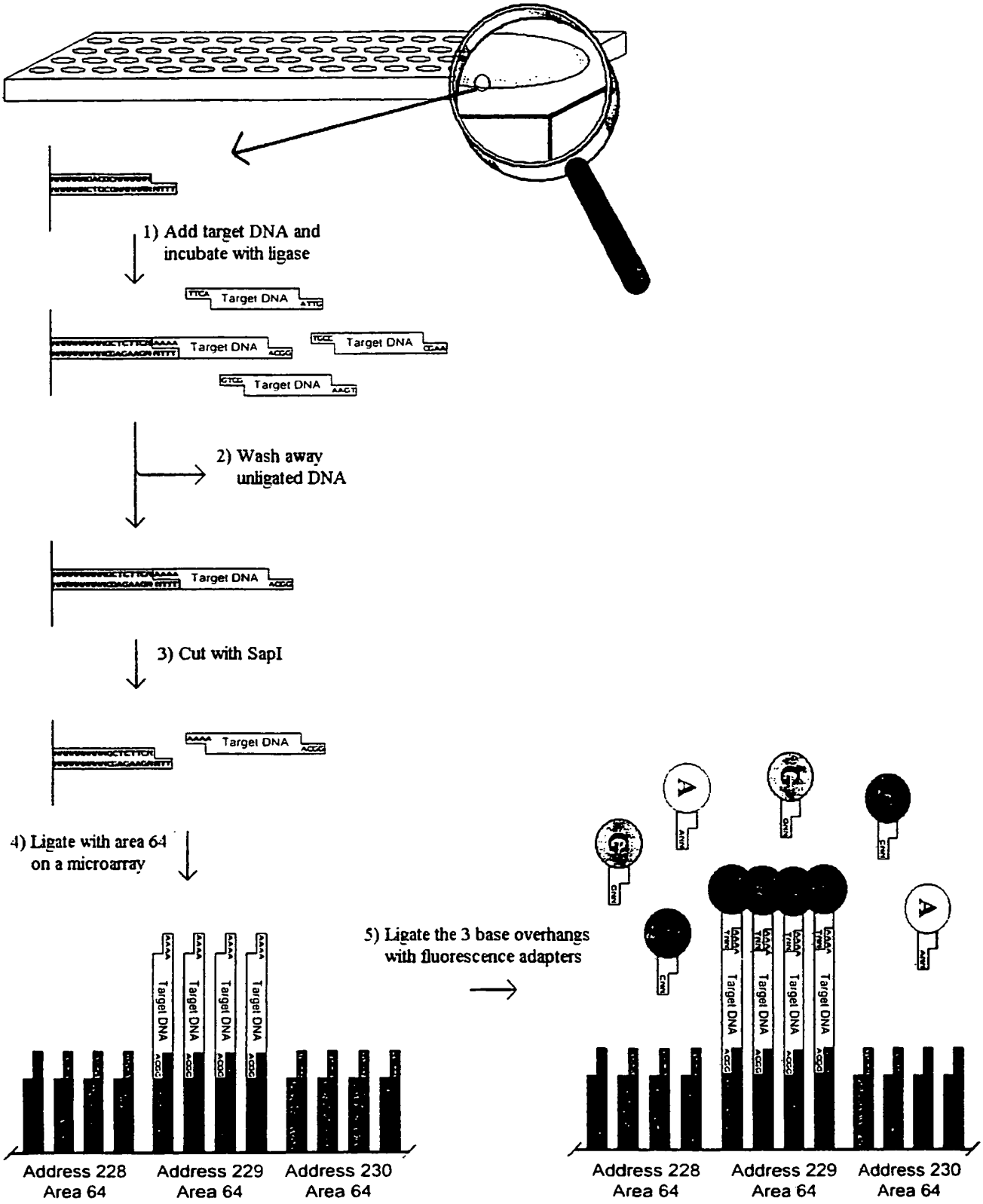


FIG. 2

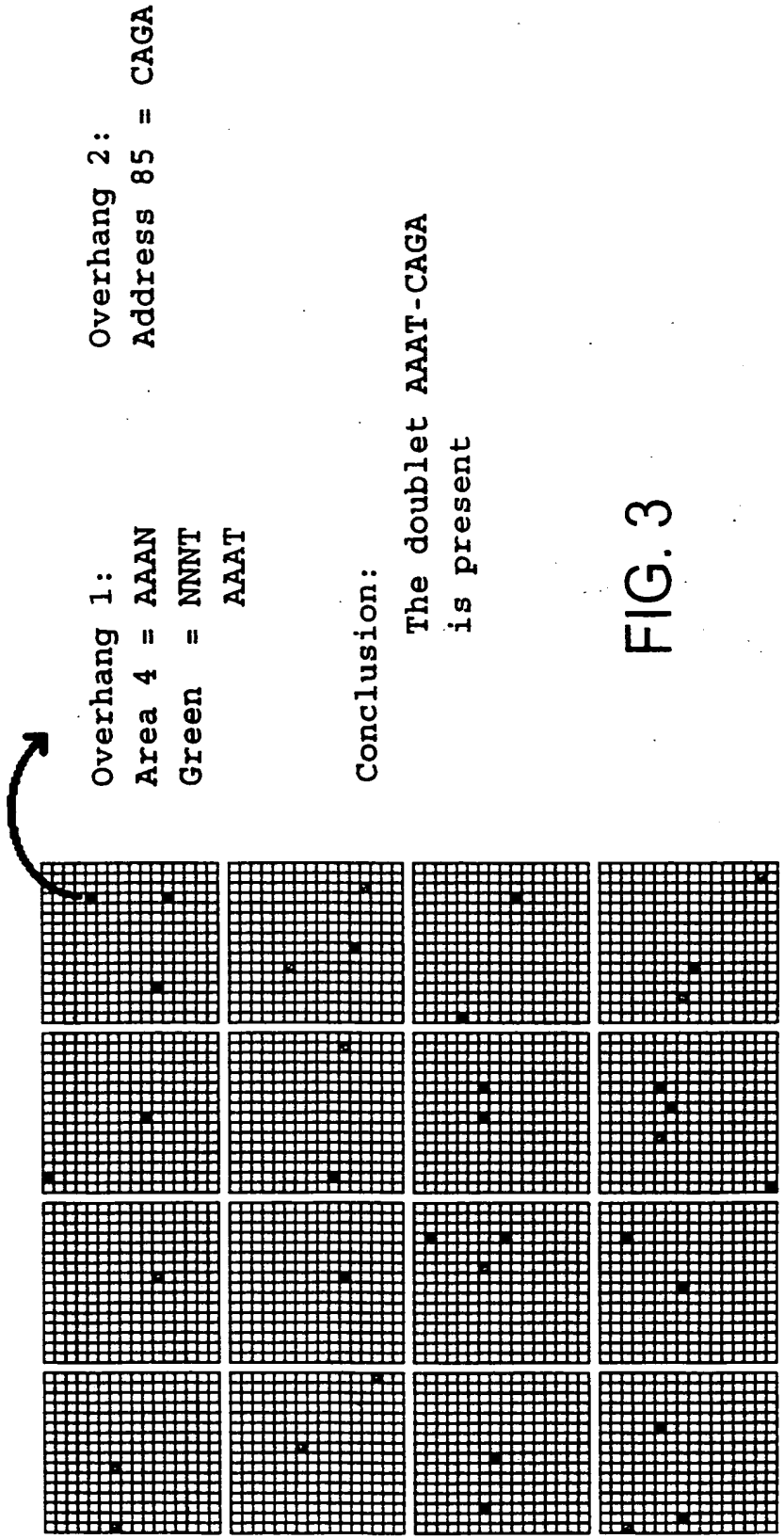

Overhang 1:
Area 4 = AAAN
Green = NNNT
AAAT
Overhang 2:
Address 85 = CAGA
Conclusion:
The doublet AAAT-CAGA
is present

FIG. 3

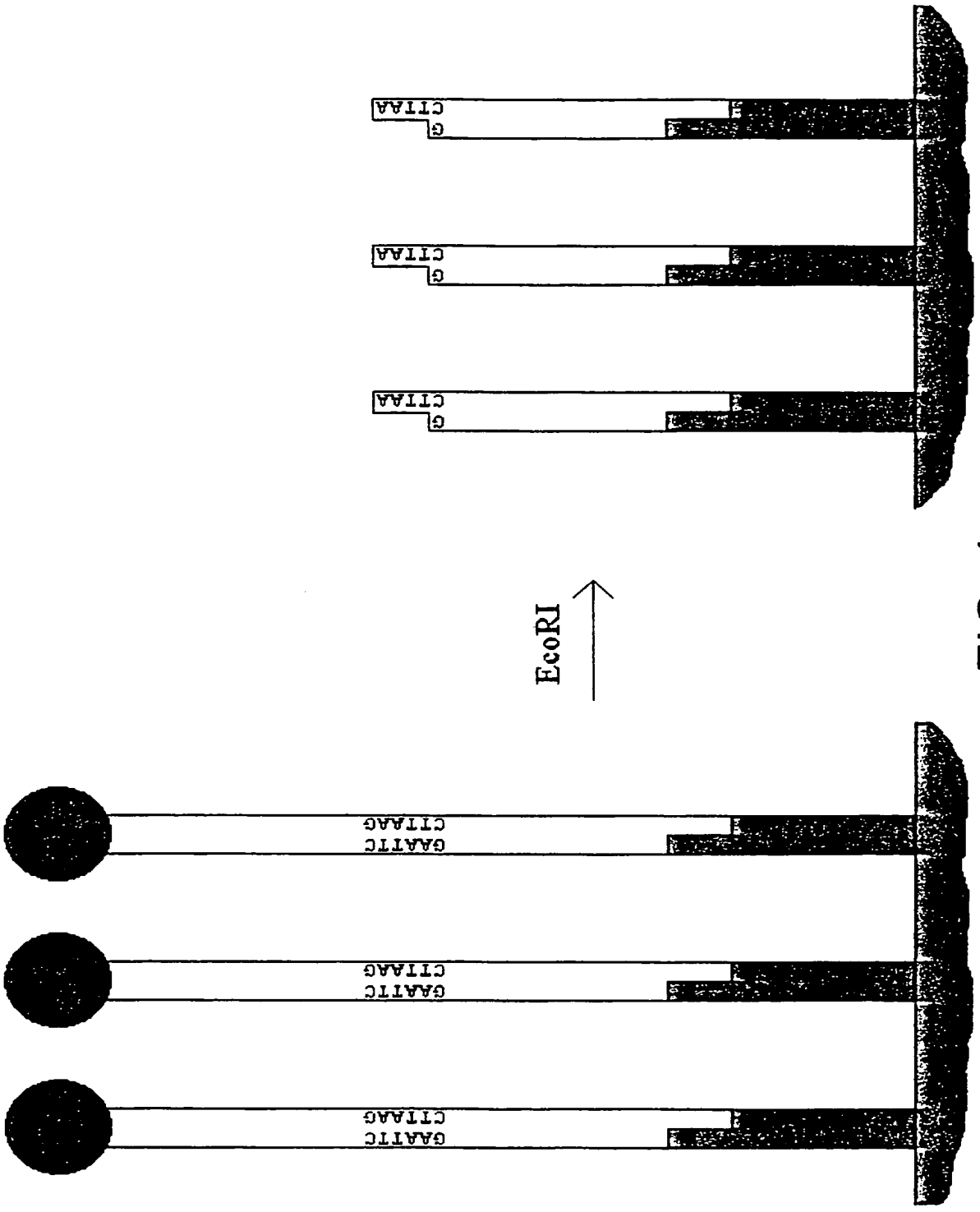

CTTAA
G
CTTAA
G
CTTAA
G
EcoRI
GAATTC
CTTAAG
GAATTC
CTTAAG
GAATTC
CTTAAG

FIG. 4

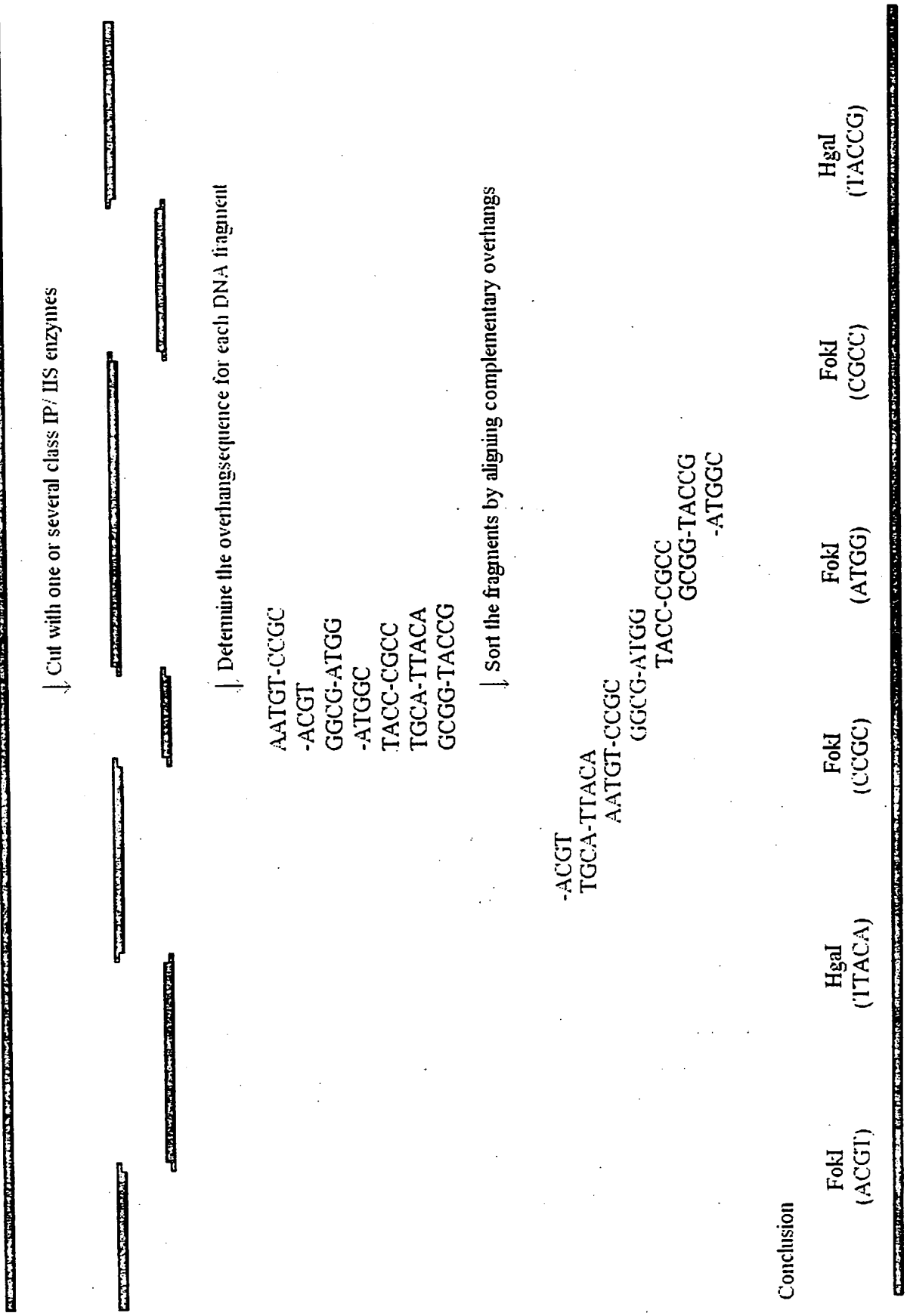
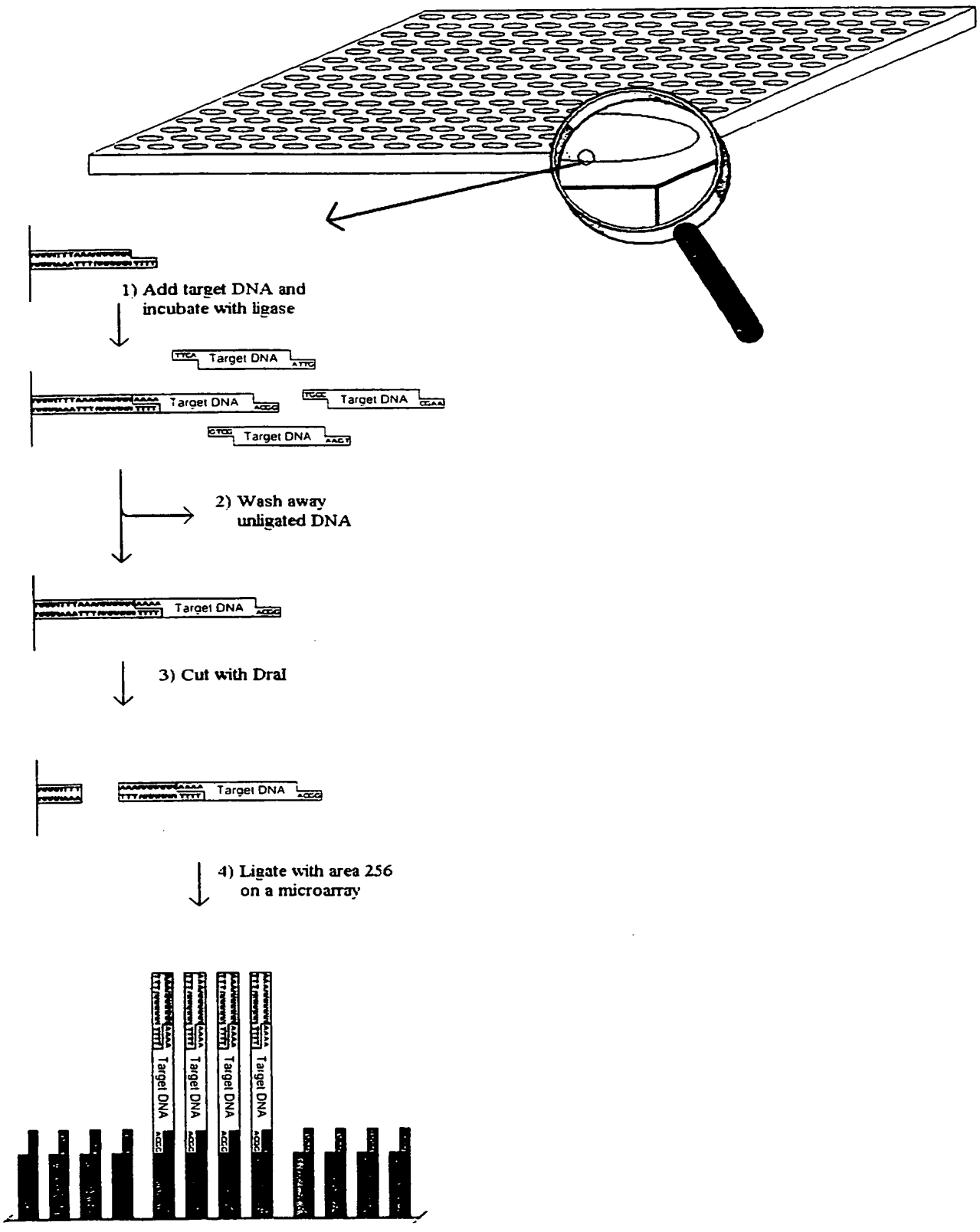


FIG. 5



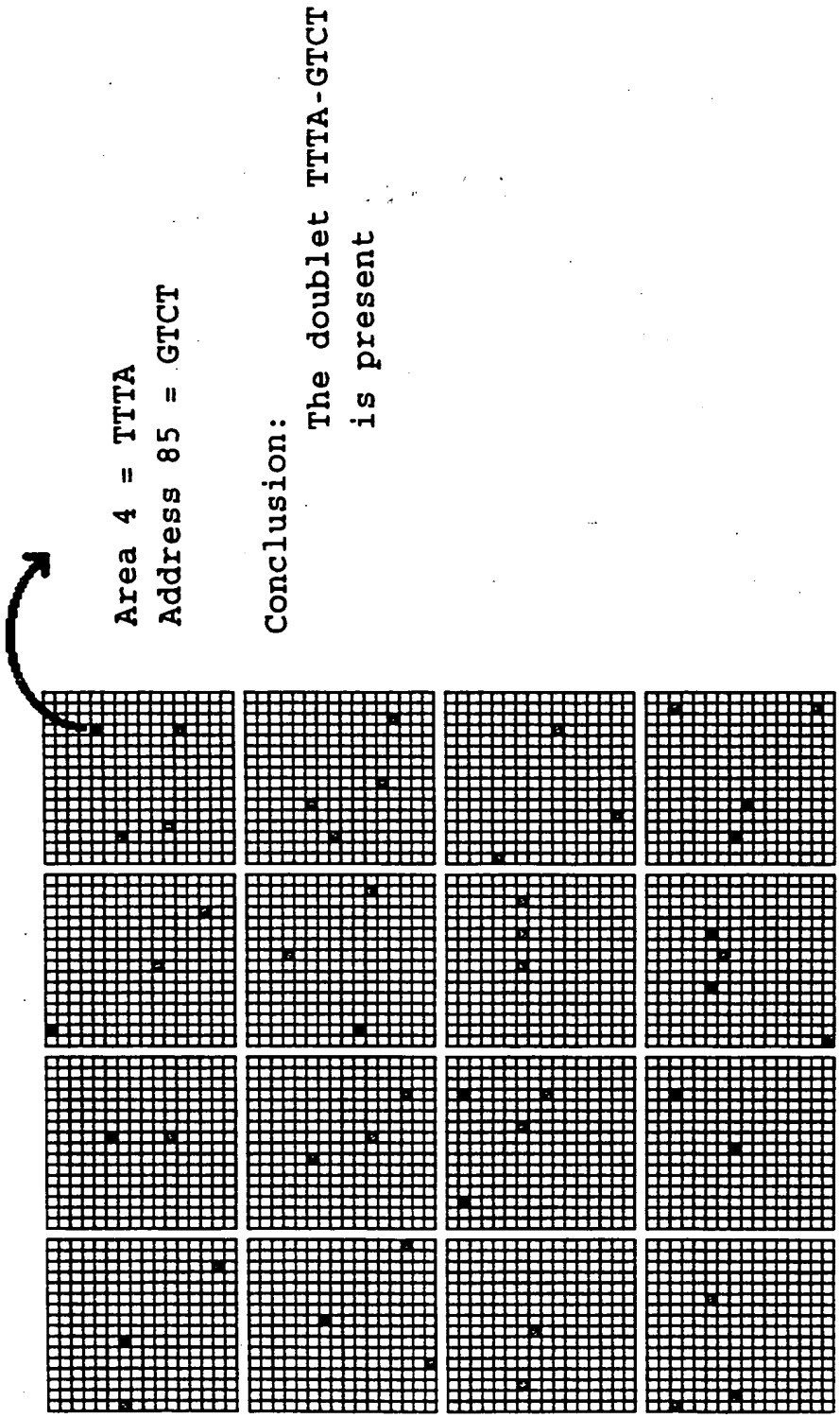


FIG. 7

(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
23 August 2001 (23.08.2001)

PCT

(10) International Publication Number
WO 01/061036 A3

(51) International Patent Classification[7]: C12Q 1/68

(21) International Application Number: PCT/GB01/00718

(22) International Filing Date: 19 February 2001 (19.02.2001)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:

20000792	17 February 2000 (17.02.2000)	NO
20012864	21 February 2000 (21.02.2000)	NO
20012863	27 February 2000 (27.02.2000)	NO

(71) Applicant *(for all designated States except US)*: **COMPLETE GENOMICS AS** [NO/NO]; P.O. Box 64, Blindern, N-0313 Oslo (NO).

(71) Applicant *(for MW only)*: **TOWLER, Philip, Dean** [GB/GB]; Frank B. Dehn & Co., 179 Queen Victoria Street, London EC4V 4EL (GB).

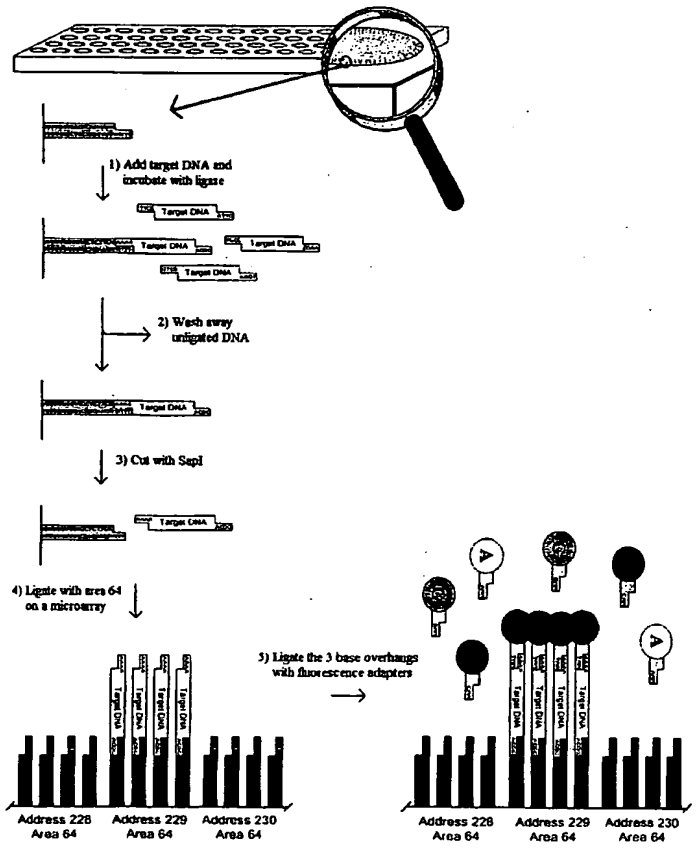
(72) Inventor; and
(75) Inventor/Applicant *(for US only)*: **LEXOW, Preben** [NO/NO]; Bloksbergveien 16, N-3132 Husøysund (NO).

(74) Agents: **TOWLER, Philip, Dean** et al.; Frank B. Dehn & Co., 179 Queen Victoria Street, London EC4V 4EL (GB).

(81) Designated States *(national)*: AE, AG, AL, AM, AT (utility model), AU, AZ, BA, BB, BG, BR, BY, BZ, CA, CH, CN, CR, CU, CZ (utility model), DE (utility model), DK (utility model), DM, DZ, EE (utility model), ES, FI (utility model), GB, GD, GE, GH, GM, HR, HU, ID, IL, IN, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MA, MD, MG, MK, MN, MW, MX, MZ, NO, NZ, PL, PT, RO, RU, SD, SE, SG, SI, SK (utility model), SL, TJ, TM, TR, TT, TZ, UA, UG, US, UZ, VN, YU, ZA, ZW.

[Continued on next page]

(54) Title: A METHOD OF MAPPING RESTRICTION ENDONUCLEASE CLEAVAGE SITES



(57) Abstract: The invention provides a two-step sorting procedure where it is possible to scan the overhanging single-stranded ends of nucleic acid fragments quickly and efficiently using solid supports, such as microarrays. Use is made of two different sets of degenerate overhang-adaptors in this regard. The invention also provides new methods and strategies inter alia for collecting information about sequences and cleavage sites that are between the cleavage sites that have generated an overhang pair. An effective method of producing the restriction map, making it easier to create multiple maps, is also described.

WO 01/061036 A3

(84) **Designated States** *(regional)*: ARIPO patent (GH, GM, KE, LS, MW, MZ, SD, SL, SZ, TZ, UG, ZW), Eurasian patent (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European patent (AT, BE, CH, CY, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE, TR), OAPI patent (BF, BJ, CF, CG, CI, CM, GA, GN, GW, ML, MR, NE, SN, TD, TG).

Published:
— *with international search report*

(88) **Date of publication of the international search report:**
12 September 2002

For two-letter codes and other abbreviations, refer to the "Guidance Notes on Codes and Abbreviations" appearing at the beginning of each regular issue of the PCT Gazette.

INTERNATIONAL SEARCH REPORT

International Application No
PCT/GB 01/00718

A. CLASSIFICATION OF SUBJECT MATTER
IPC 7 C12Q1/68

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)
IPC 7 C12Q

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practical, search terms used)
EPO-Internal, WPI Data, PAJ, MEDLINE, BIOSIS

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category °	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	US 5 858 656 A (DEUGAU KENNETH VICTOR ET AL) 12 January 1999 (1999-01-12) column 8-10; claims	1-8,14, 15,19,20
X	WO 98 10095 A (BRAX GENOMICS LTD ;THOMPSON ANDREW HUGIN (GB); SCHMIDT GUNTER (GB)) 12 March 1998 (1998-03-12) the whole document	1-8,19, 20
X	WO 99 28505 A (ROTHBERG JONATHAN M ;CURAGEN CORP (US); HU XINGHUA (US); NALLUR GI) 10 June 1999 (1999-06-10) page 8-9; figures 6,7 page 89-90; figures 6,7	9,10

-/--

[X] Further documents are listed in the continuation of box C.

[X] Patent family members are listed in annex.

° Special categories of cited documents :

"A" document defining the general state of the art which is not considered to be of particular relevance

"E" earlier document but published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art.

"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
27 May 2002	04/06/2002

Name and mailing address of the ISA	Authorized officer
European Patent Office, P.B. 5818 Patentlaan 2 NL – 2280 HV Rijswijk Tel. (+31–70) 340–2040, Tx. 31 651 epo nl, Fax: (+31–70) 340–3016	Reuter, U

Form PCT/ISA/210 (second sheet) (July 1992)

INTERNATIONAL SEARCH REPORT

International Application No
PCT/GB 01/00718

C.(Continuation) DOCUMENTS CONSIDERED TO BE RELEVANT

Category °	Citation of document, with indication,where appropriate, of the relevant passages	Relevant to claim No.
A	WO 99 00519 A (BRENNER SYDNEY ;LYNX THERAPEUTICS INC (US)) 7 January 1999 (1999-01-07) the whole document	1-20
A	US 5 728 524 A (SIBSON DAVID ROSS) 17 March 1998 (1998-03-17) the whole document	1-20

INTERNATIONAL SEARCH REPORT

Information on patent family members

International Application No
PCT/GB 01/00718

Patent document cited in search report		Publication date	Patent family member(s)		Publication date
US 5858656	A	12-01-1999	US	5508169 A	16-04-1996
			CA	2036946 A1	07-10-1991
WO 9810095	A	12-03-1998	AT	215994 T	15-04-2002
			AU	721861 B2	13-07-2000
			AU	4027497 A	26-03-1998
			CN	1234076 A	03-11-1999
			DE	69711895 D1	16-05-2002
			EP	0927267 A1	07-07-1999
			WO	9810095 A1	12-03-1998
			JP	2000517192 T	26-12-2000
			NZ	334426 A	29-09-2000
			US	6225077 B1	01-05-2001
WO 9928505	A	10-06-1999	AU	1603199 A	16-06-1999
			WO	9928505 A1	10-06-1999
			US	6355423 B1	12-03-2002
WO 9900519	A	07-01-1999	AU	739963 B2	25-10-2001
			AU	8170998 A	19-01-1999
			EP	1017847 A1	12-07-2000
			JP	2002507126 T	05-03-2002
			WO	9900519 A1	07-01-1999
US 5728524	A	17-03-1998	AT	159986 T	15-11-1997
			AU	686563 B2	12-02-1998
			AU	4575893 A	31-01-1994
			CA	2139944 A1	20-01-1994
			DE	69315074 D1	11-12-1997
			DE	69315074 T2	05-03-1998
			EP	0650528 A1	03-05-1995
			WO	9401582 A1	20-01-1994
			JP	7508883 T	05-10-1995
			JP	3251291 B2	28-01-2002

(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

CORRECTED VERSION

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
23 August 2001 (23.08.2001)

PCT

(10) International Publication Number
WO 2001/061036 A3

(51) International Patent Classification[7]: C12Q 1/68

(21) International Application Number:
PCT/GB2001/000718

(22) International Filing Date: 19 February 2001 (19.02.2001)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:

20000792	17 February 2000 (17.02.2000)	NO
20012864	21 February 2000 (21.02.2000)	NO
20012863	27 February 2000 (27.02.2000)	NO

(71) Applicant *(for all designated States except US)*: **COMPLETE GENOMICS AS** [NO/NO]; P.O. Box 64, Blindern, N-0313 Oslo (NO).

(71) Applicant *(for MW only)*: **TOWLER, Philip, Dean** [GB/GB]; Frank B. Dehn & Co., 179 Queen Victoria Street, London EC4V 4EL (GB).

(72) Inventor; and

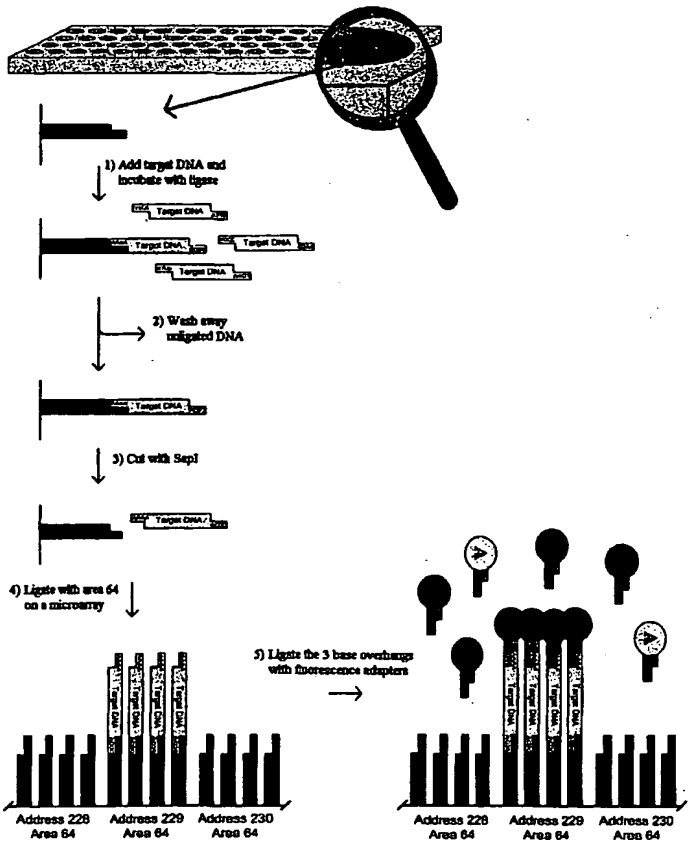
(75) Inventor/Applicant *(for US only)*: **LEXOW, Preben** [NO/NO]; Bloksbergveien 16, N-3132 Husøysund (NO).

(74) Agents: **TOWLER, Philip, Dean** et al.; Frank B. Dehn & Co., 179 Queen Victoria Street, London EC4V 4EL (GB).

(81) Designated States *(national)*: AE, AG, AL, AM, AT (utility model), AT, AU, AZ, BA, BB, BG, BR, BY, BZ, CA, CH, CN, CR, CU, CZ (utility model), CZ, DE (utility model), DE, DK (utility model), DK, DM, DZ, EE (utility model), EE, ES, FI (utility model), FI, GB, GD, GE, GH, GM, HR, HU, ID, IL, IN, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MA, MD, MG, MK, MN,

[Continued on next page]

(54) Title: A METHOD OF MAPPING RESTRICTION ENDONUCLEASE CLEAVAGE SITES



(57) Abstract: The invention provides a two-step sorting procedure where it is possible to scan the overhanging single-stranded ends of nucleic acid fragments quickly and efficiently using solid supports, such as microarrays. Use is made of two different sets of degenerate overhang-adaptors in this regard. The invention also provides new methods and strategies inter alia for collecting information about sequences and cleavage sites that are between the cleavage sites that have generated an overhang pair. An effective method of producing the restriction map, making it easier to create multiple maps, is also described.

WO 2001/061036 A3

MW, MX, MZ, NO, NZ, PL, PT, RO, RU, SD, SE, SG, SI, SK (utility model), SK, SL, TJ, TM, TR, TT, TZ, UA, UG, US, UZ, VN, YU, ZA, ZW.

(84) **Designated States** *(regional)*: ARIPO patent (GH, GM, KE, LS, MW, MZ, SD, SL, SZ, TZ, UG, ZW), Eurasian patent (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European patent (AT, BE, CH, CY, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE, TR), OAPI patent (BF, BJ, CF, CG, CI, CM, GA, GN, GW, ML, MR, NE, SN, TD, TG).

Published:

— *with international search report*

(88) **Date of publication of the international search report:**
12 September 2002

(48) **Date of publication of this corrected version:**
15 April 2004

(15) **Information about Correction:**
see PCT Gazette No. 16/2004 of 15 April 2004, Section II

For two-letter codes and other abbreviations, refer to the "Guidance Notes on Codes and Abbreviations" appearing at the beginning of each regular issue of the PCT Gazette.

This Page is inserted by IFW Indexing and Scanning Operations and is not part of the Official Record

BEST AVAILABLE IMAGES

Defective images within this document are accurate representations of the original documents submitted by the applicant.

Defects in the images include but are not limited to the items checked:

- ☒ BLACK BORDERS
- ☒ IMAGE CUT OFF AT TOP, BOTTOM OR SIDES
- ☒ FADED TEXT OR DRAWING
- ☒ BLURED OR ILLEGIBLE TEXT OR DRAWING
- ☒ SKEWED/SLANTED IMAGES
- ☐ COLORED OR BLACK AND WHITE PHOTOGRAPHS
- ☐ GRAY SCALE DOCUMENTS
- ☐ LINES OR MARKS ON ORIGINAL DOCUMENT
- ☐ REPERENCE(S) OR EXHIBIT(S) SUBMITTED ARE POOR QUALITY
- ☐ OTHER: _____

IMAGES ARE BEST AVAILABLE COPY.
As rescanning documents *will not* correct images problems checked, please do not report the problems to the IFW Image Problem Mailbox